Manifold Sampling for Nonconvex Optimization of Piecewise Linear Compositions

Kamil Khan, Jeffrey Larson, Matt Menickelly, Stefan Wild

Argonne National Laboratory

July 4, 2018

Problem statement

We are interested in solving the problem:

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x) \triangleq \psi(x) + h(F(x))$$

where $\psi : \mathbb{R}^n \to \mathbb{R}$, $F : \mathbb{R}^n \to \mathbb{R}^p$, $h : \mathbb{R}^p \to \mathbb{R}$,

Problem statement

We are interested in solving the problem:

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x) \triangleq \psi(x) + h(F(x))$$

where $\psi : \mathbb{R}^n \to \mathbb{R}$, $F : \mathbb{R}^n \to \mathbb{R}^p$, $h : \mathbb{R}^p \to \mathbb{R}$, and

- ψ is smooth with known derivatives

Problem statement

We are interested in solving the problem:

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x) \triangleq \psi(x) + h(F(x))$$

where $\psi : \mathbb{R}^n \to \mathbb{R}$, $F : \mathbb{R}^n \to \mathbb{R}^p$, $h : \mathbb{R}^p \to \mathbb{R}$, and

- ψ is smooth with known derivatives
- h is nonsmooth, piecewise linear, and has a known structure (cheap to evaluate)

Problem statement

We are interested in solving the problem:

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x) \triangleq \psi(x) + h(F(x))$$

where $\psi : \mathbb{R}^n \to \mathbb{R}$, $F : \mathbb{R}^n \to \mathbb{R}^p$, $h : \mathbb{R}^p \to \mathbb{R}$, and

- ψ is smooth with known derivatives
- h is nonsmooth, piecewise linear, and has a known structure (cheap to evaluate)
- F is smooth, nonlinear, and has a relatively unknown structure (expensive to evaluate)

Problem statement

We are interested in solving the problem:

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x) \triangleq \psi(x) + h(F(x))$$

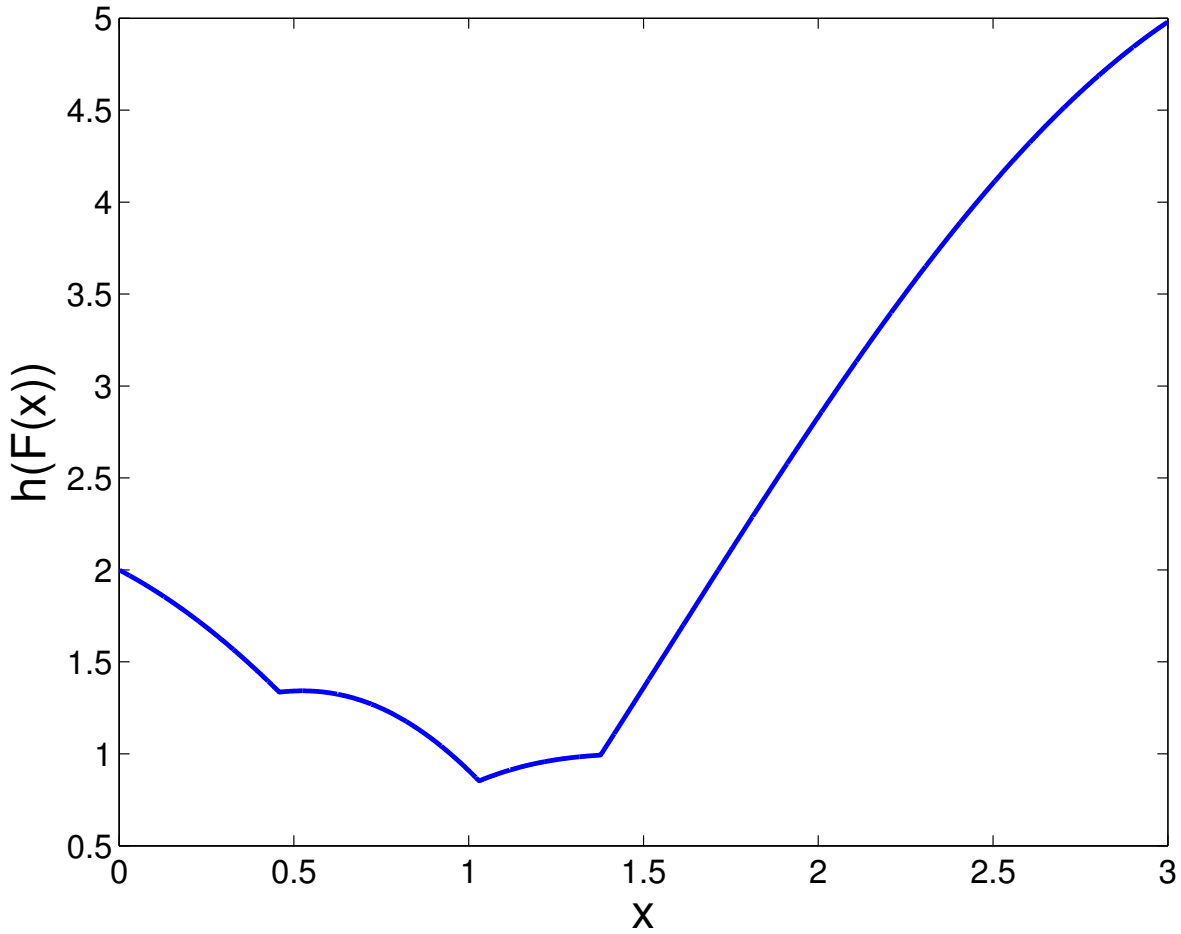
where $\psi : \mathbb{R}^n \to \mathbb{R}$, $F : \mathbb{R}^n \to \mathbb{R}^p$, $h : \mathbb{R}^p \to \mathbb{R}$, and

- ψ is smooth with known derivatives
- h is nonsmooth, piecewise linear, and has a known structure (cheap to evaluate)
- F is smooth, nonlinear, and has a relatively unknown structure (expensive to evaluate)

Piecewise linear h does not imply $h \circ F$ is piecewise linear.

Formulation

$$h(F(x)) = \max\{\sin(2x) + 1, \cos(2x), x\} - \min\{\sin(2x) + 1, \cos(2x), x\}$$



Notes

- The *manifold sampling* framework does not require the availability of the Jacobian ∇F.

Notes

- The *manifold sampling* framework does not require the availability of the Jacobian ∇F.

- Applicable both when inexact values for $\nabla F(x)$ are available and in the derivative-free case, when only $F(x)$ is available.

Notes

- The *manifold sampling* framework does not require the availability of the Jacobian ∇F.

- Applicable both when inexact values for $\nabla F(x)$ are available and in the derivative-free case, when only $F(x)$ is available.

- We will build component models m^{F_i} of each F_i around points x. We can then use $\nabla M(x) \in \mathbb{R}^{n \times p}$ where

$$\nabla M(x) \triangleq \left[\nabla m^{F_1}(x), \ldots, \nabla m^{F_p}(x)\right].$$

Piecewise linear functions

Definition

A function $h\colon \mathbb{R}^p \to \mathbb{R}$ is *piecewise linear* if h is continuous and there exists a finite collection $\mathfrak{H} \triangleq \{h_i : i = 1, \ldots, \hat{m}\}$ of affine functions that map $\mathbb{R}^p$ into $\mathbb{R}$, for which

$$h(z) \in \left\{\tilde{h}(z) : \tilde{h} \in \mathfrak{H}\right\}, \qquad \forall z \in \mathbb{R}^p.$$

- h is a *continuous selection* of $\mathfrak{H}$.
- Elements of $\mathfrak{H}$ are *selection functions of* h.
- $h_i : z \in \mathbb{R}^p \mapsto \langle a_i, z\rangle + b_i$ for each i.

Piecewise linear functions

Definition

A function $h\colon \mathbb{R}^p \to \mathbb{R}$ is *piecewise linear* if h is continuous and there exists a finite collection $\mathfrak{H} \triangleq \{h_i : i = 1, \ldots, \hat{m}\}$ of affine functions that map $\mathbb{R}^p$ into $\mathbb{R}$, for which

$$h(z) \in \left\{\tilde{h}(z) : \tilde{h} \in \mathfrak{H}\right\}, \qquad \forall z \in \mathbb{R}^p.$$

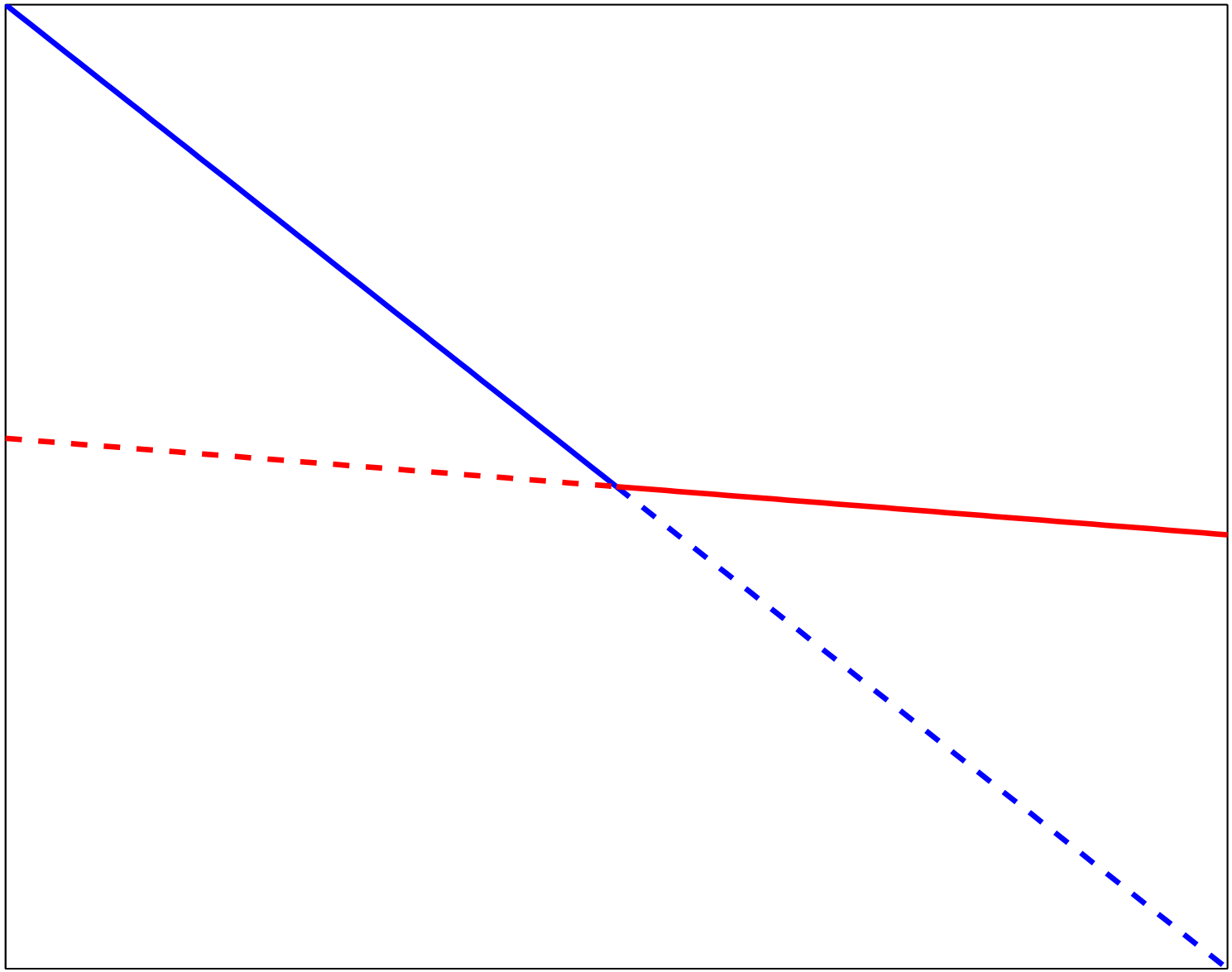
- h is a *continuous selection* of $\mathfrak{H}$.
- Elements of $\mathfrak{H}$ are *selection functions of* h.
- $h_i : z \in \mathbb{R}^p \mapsto \langle a_i, z\rangle + b_i$ for each i.

Definition

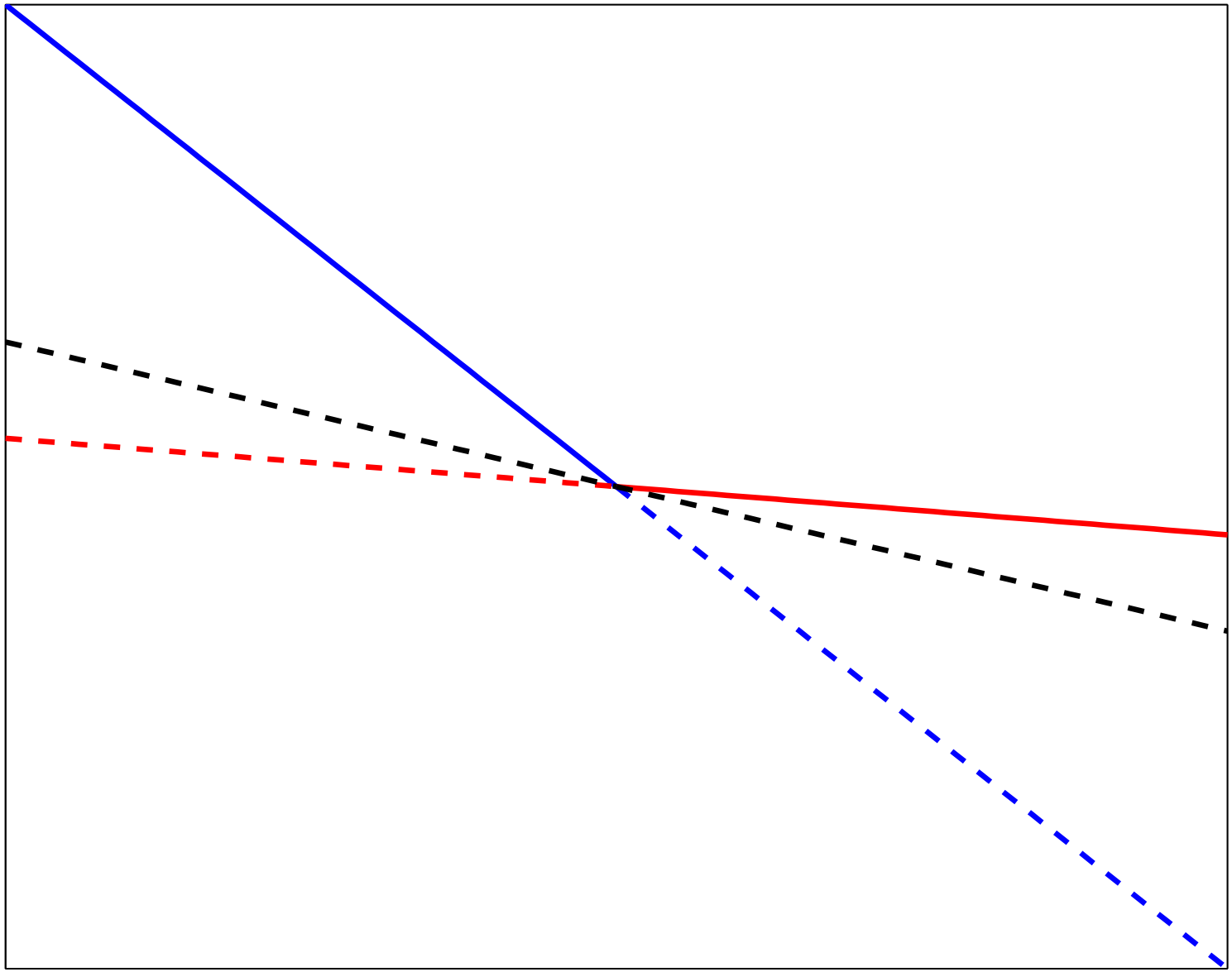
$$\mathcal{S}_i \triangleq \{y : h(y) = h_i(y)\}, \quad \tilde{\mathcal{S}}_i \triangleq \mathbf{cl}\left(\mathbf{int}\left(\mathcal{S}_i\right)\right), \quad I_h(z) \triangleq \left\{i : z \in \tilde{\mathcal{S}}_i\right\},$$

h_i for $i \in I_h(z)$ is an *essentially active selection function* for h at z.

Essentially active

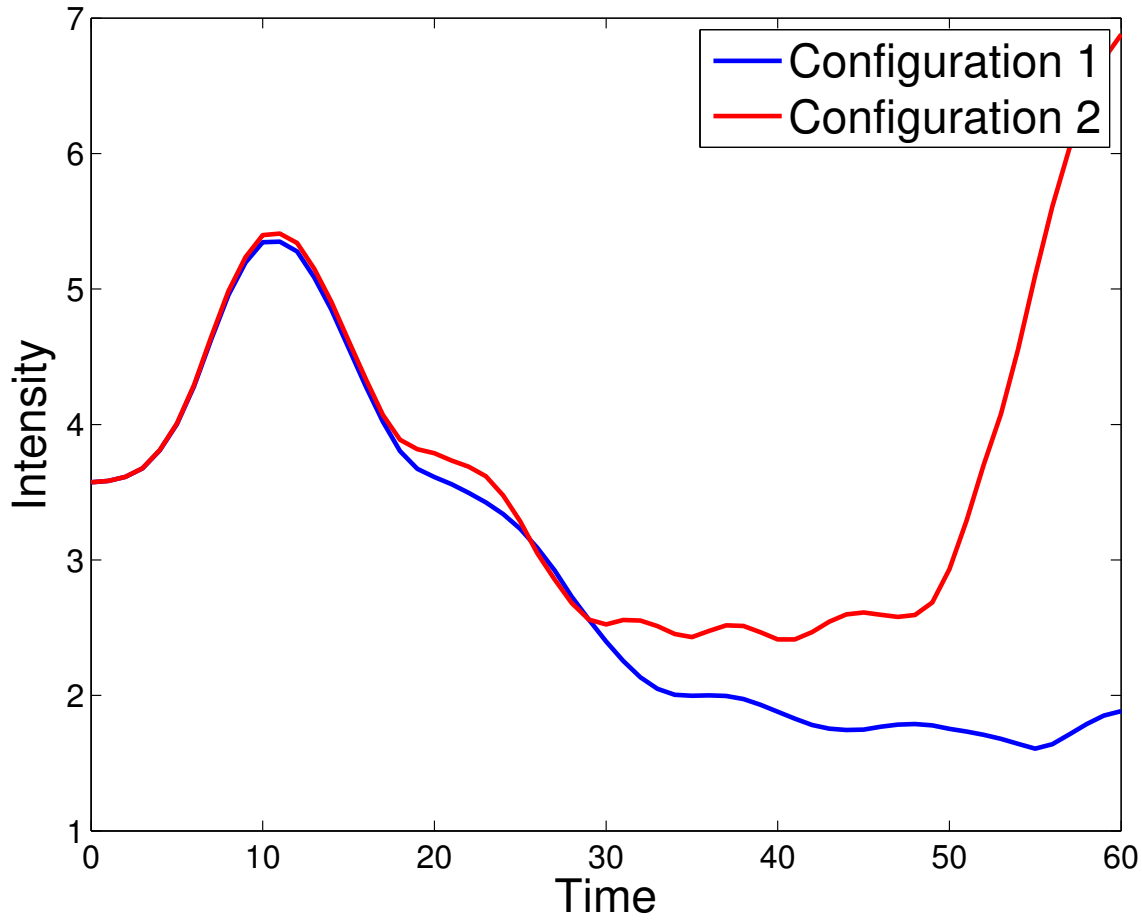


Essentially active



Laser pulse propagating in a plasma channel

Determine plasma channel properties that minimize the maximum difference in the laser intensity.



$$f(x) = \max_{\Omega_1} \{F_i(x)\} - \min_{\Omega_2} \{F_i(x)\}$$

A generalized derivative

Definition

The B-*subdifferential* of f at x is defined as

$$\partial_{\mathrm{B}} f(x) \triangleq \left\{ \xi : \xi = \lim_{y^j \to x} \nabla f(y^j) : \; y^j \in \mathcal{D} \right\}.$$

The *generalized Clarke subdifferential* of f at x is defined as

$$\partial_{\mathrm{C}} f(x) \triangleq \mathbf{co}\left(\partial_{\mathrm{B}}\right).$$

A generalized derivative

Definition

The B-*subdifferential* of f at x is defined as

$$\partial_{\mathrm{B}} f(x) \triangleq \left\{ \xi : \xi = \lim_{y^j \to x} \nabla f(y^j) : \; y^j \in \mathcal{D} \right\}.$$

The *generalized Clarke subdifferential* of f at x is defined as

$$\partial_{\mathrm{C}} f(x) \triangleq \mathbf{co}\left(\partial_{\mathrm{B}}\right).$$

For our case:

$$\partial_{\mathrm{C}} h(z) = \mathbf{co}\left(\{a_i : i \in I_h(z)\}\right)$$

A generalized derivative

Definition

The B-*subdifferential* of f at x is defined as

$$\partial_{\mathrm{B}} f(x) \triangleq \left\{ \xi : \xi = \lim_{y^j \to x} \nabla f(y^j) : \; y^j \in \mathcal{D} \right\}.$$

The *generalized Clarke subdifferential* of f at x is defined as

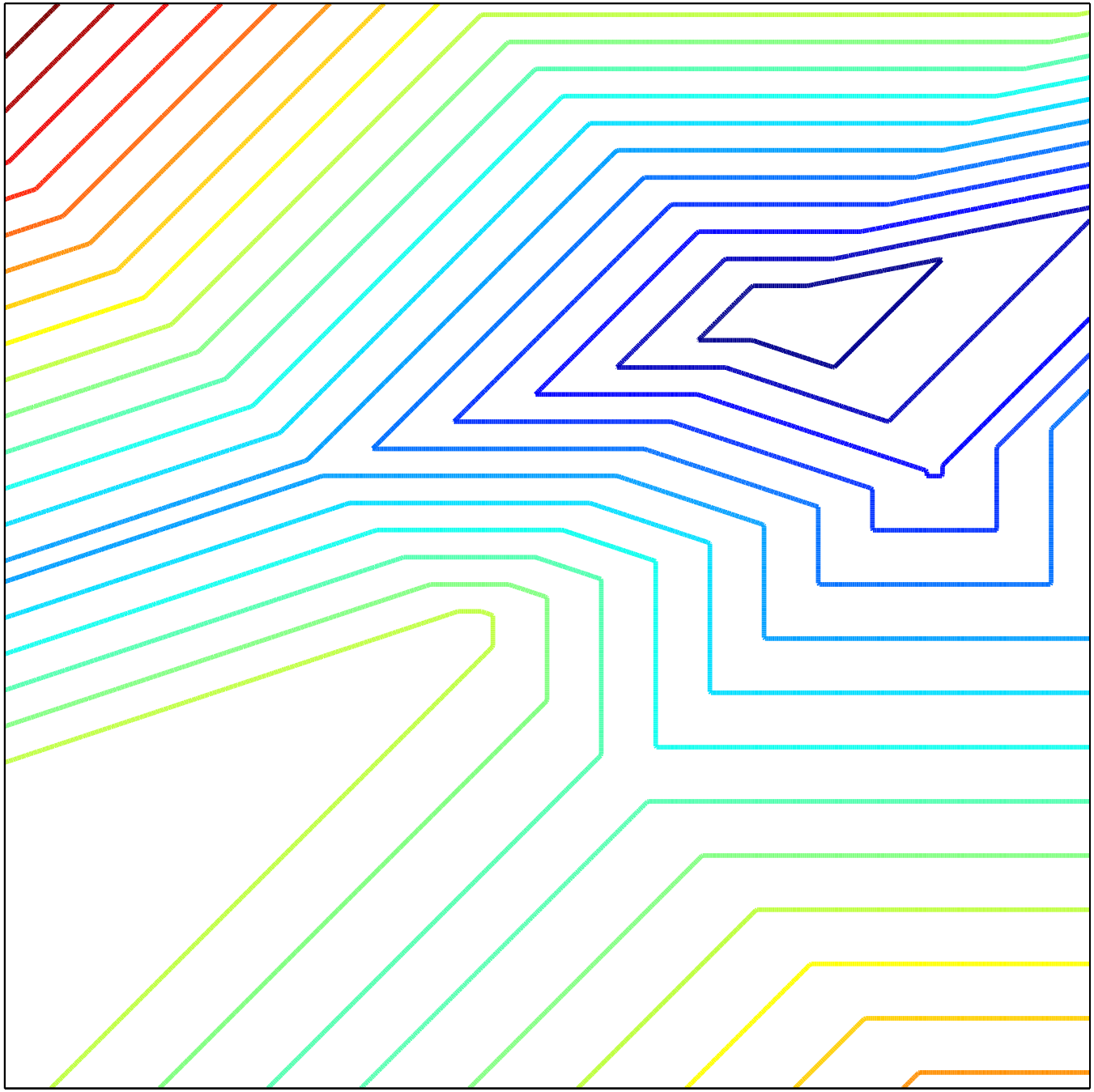
$$\partial_{\mathrm{C}} f(x) \triangleq \mathbf{co}\left(\partial_{\mathrm{B}}\right).$$

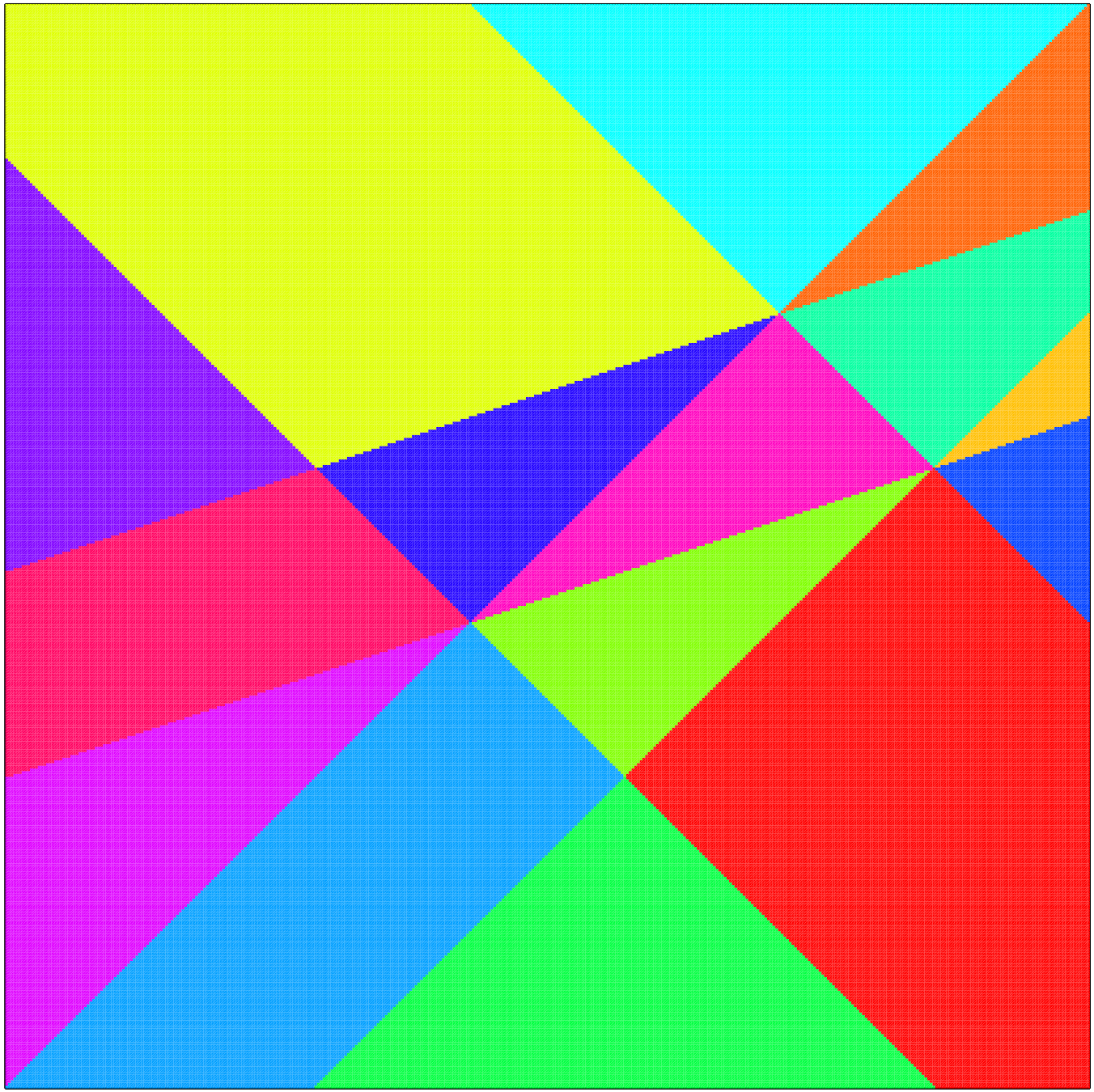
For our case:

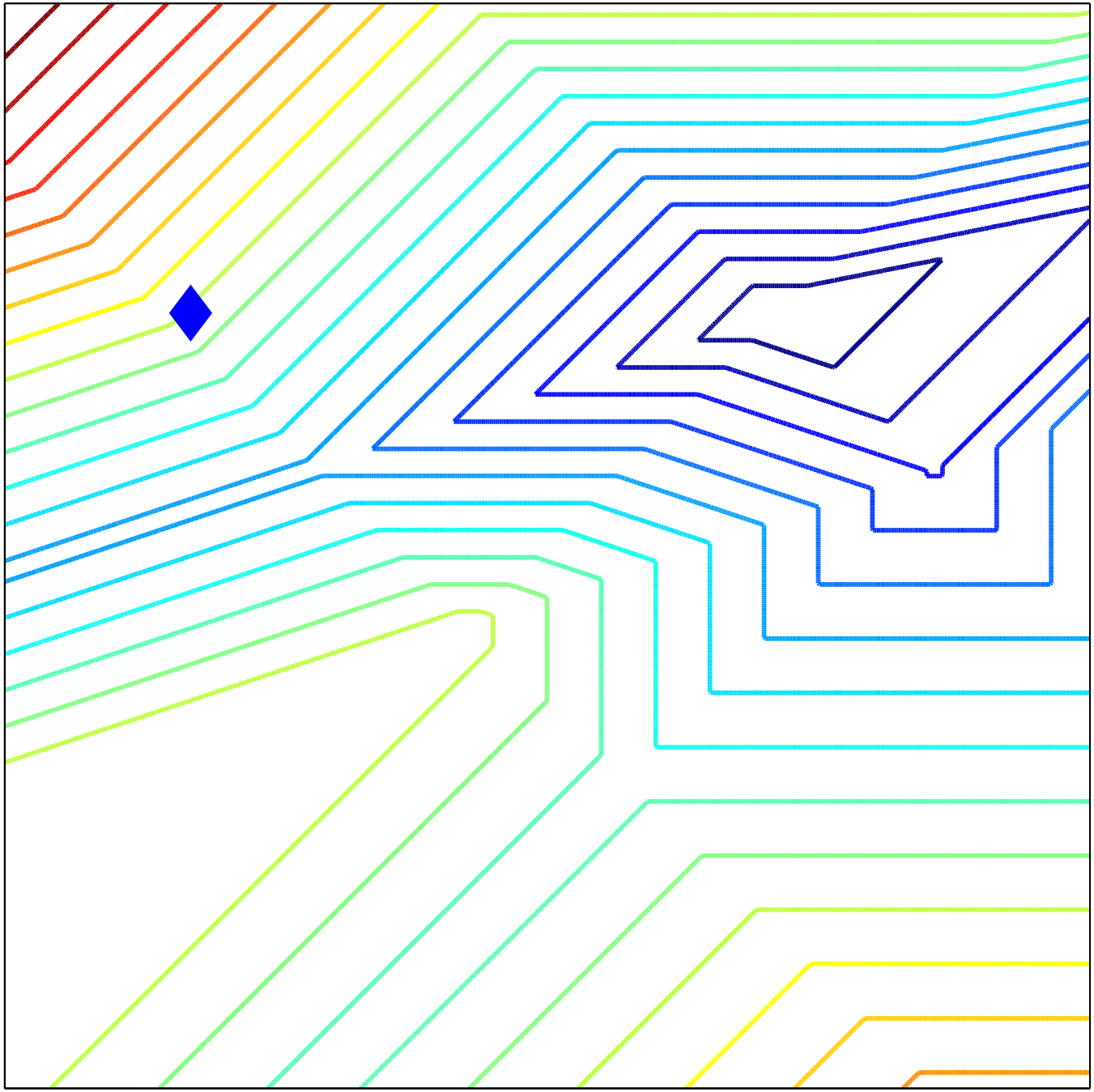
$$\partial_{\mathrm{C}} h(z) = \mathbf{co}\left(\{a_i : i \in I_h(z)\}\right)$$

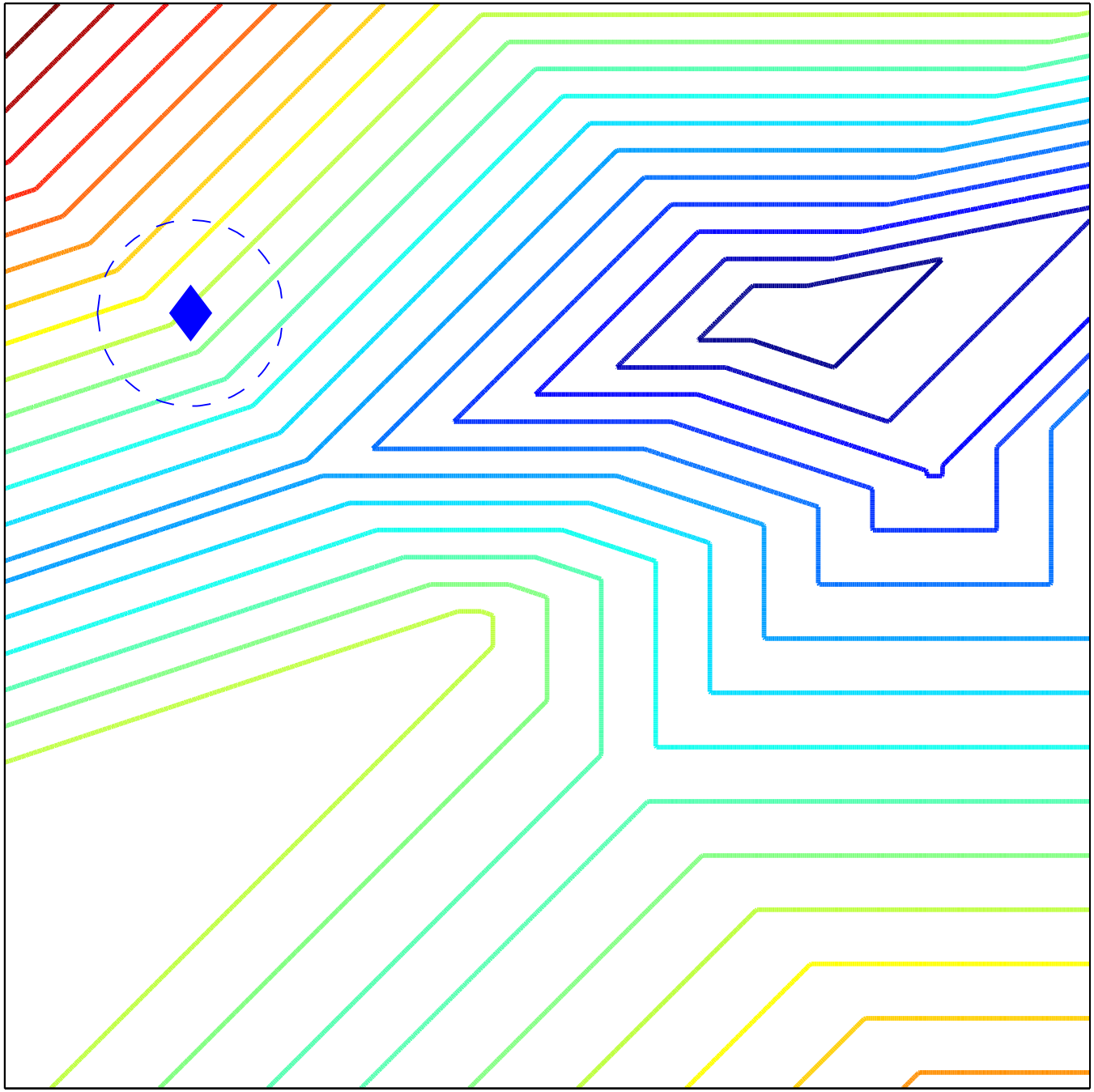
Definition

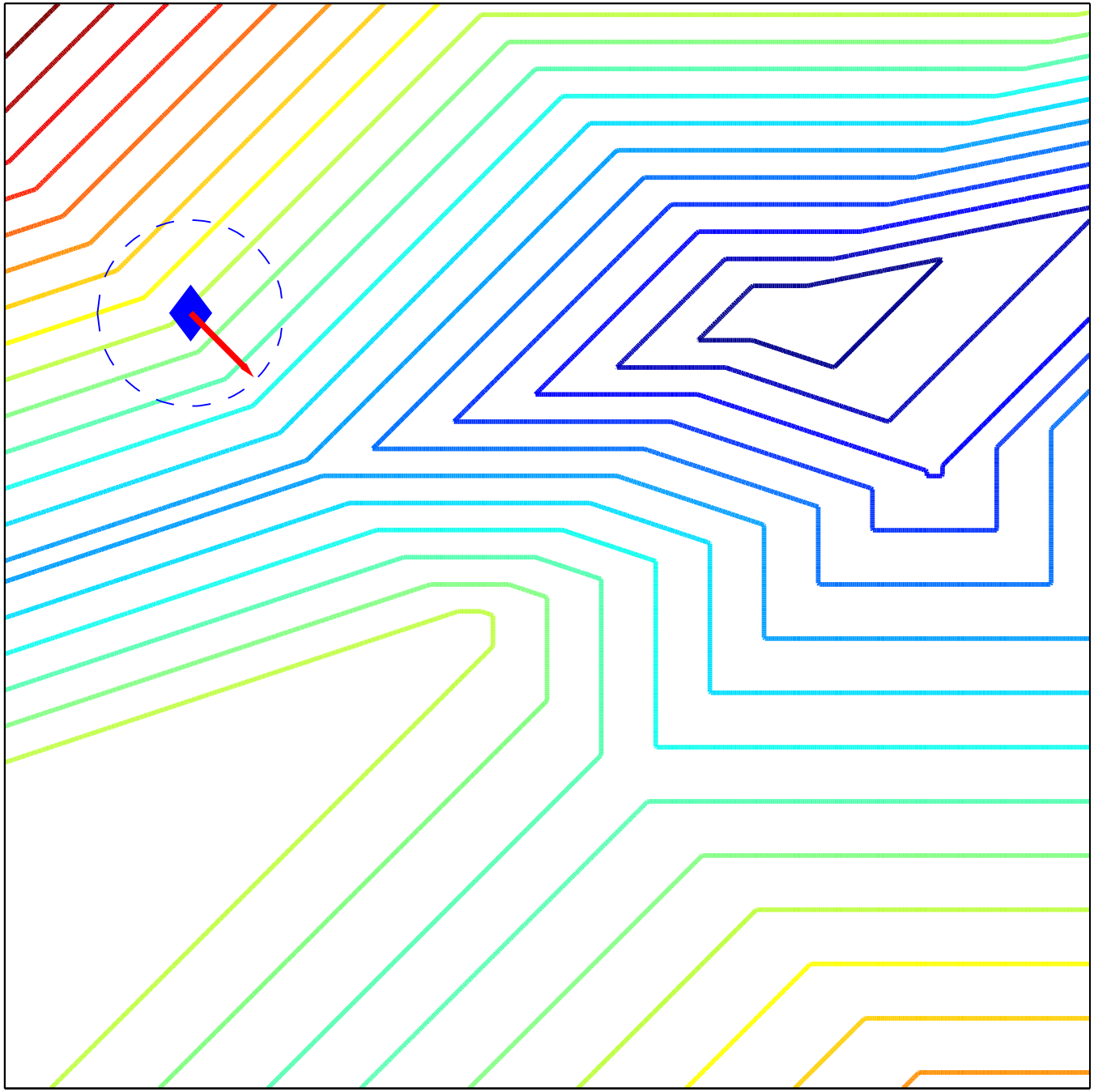
A point x is called a *Clarke stationary* point of f if $0 \in \partial_{\mathrm{C}} f(x)$.

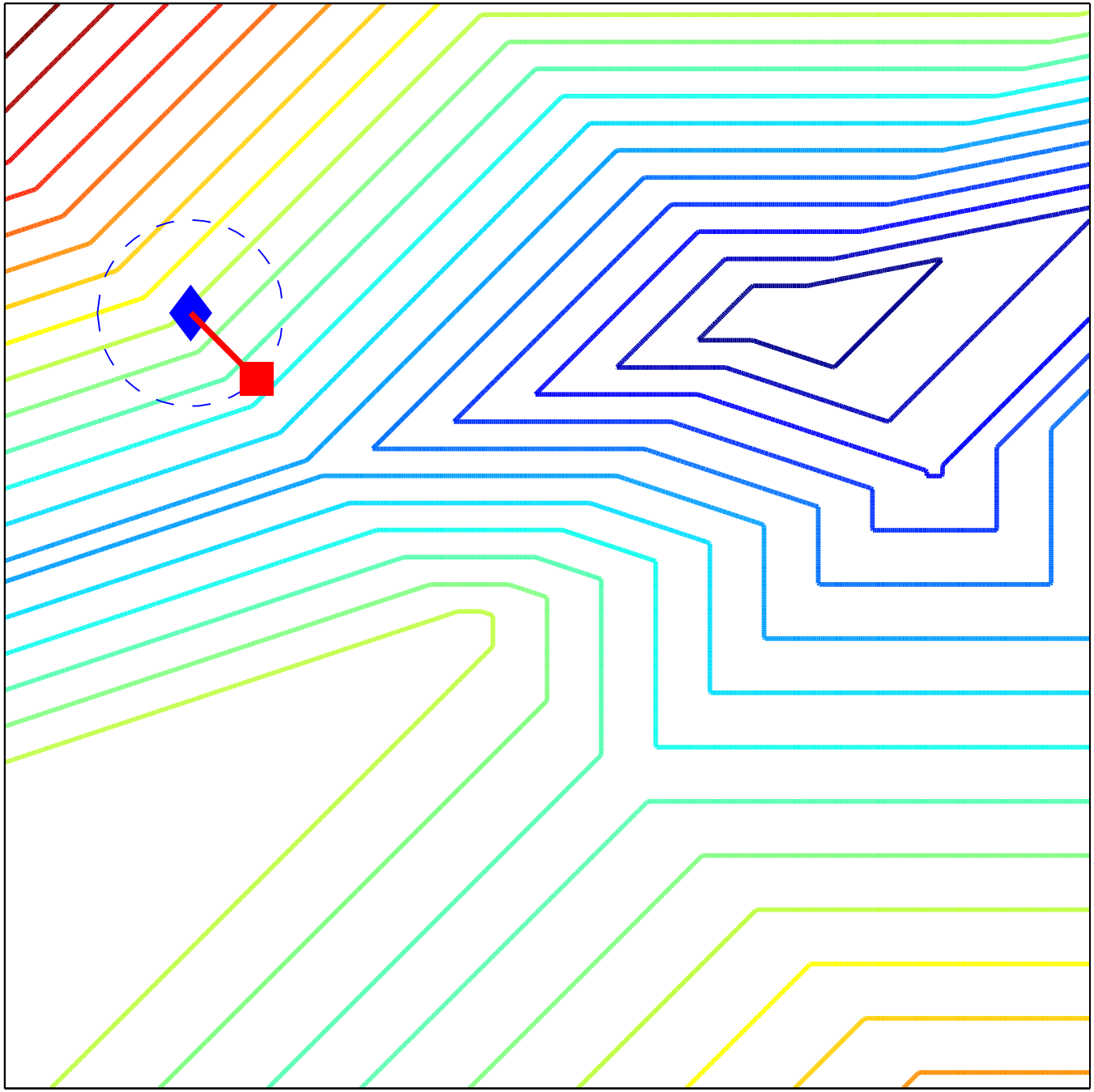


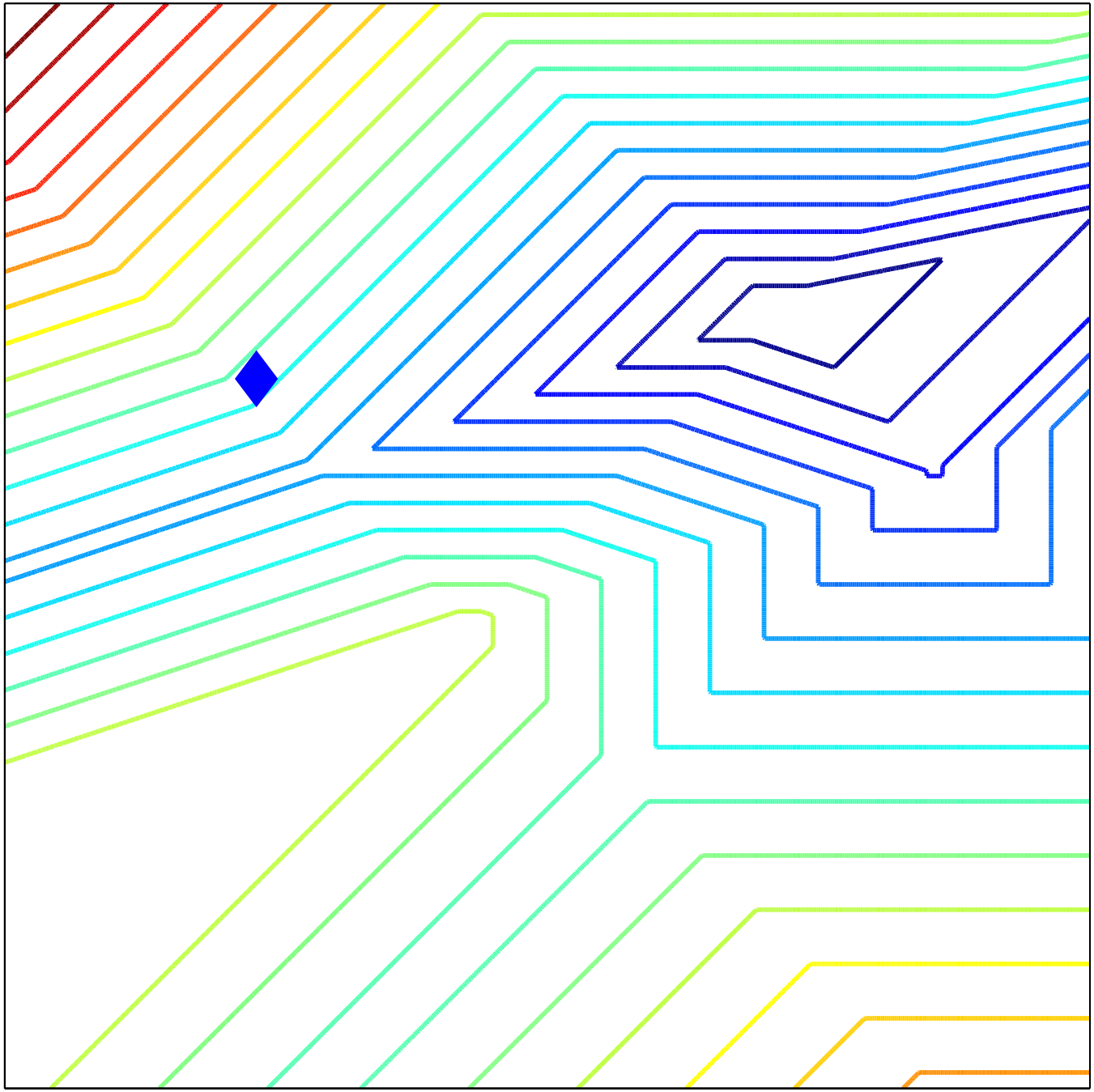


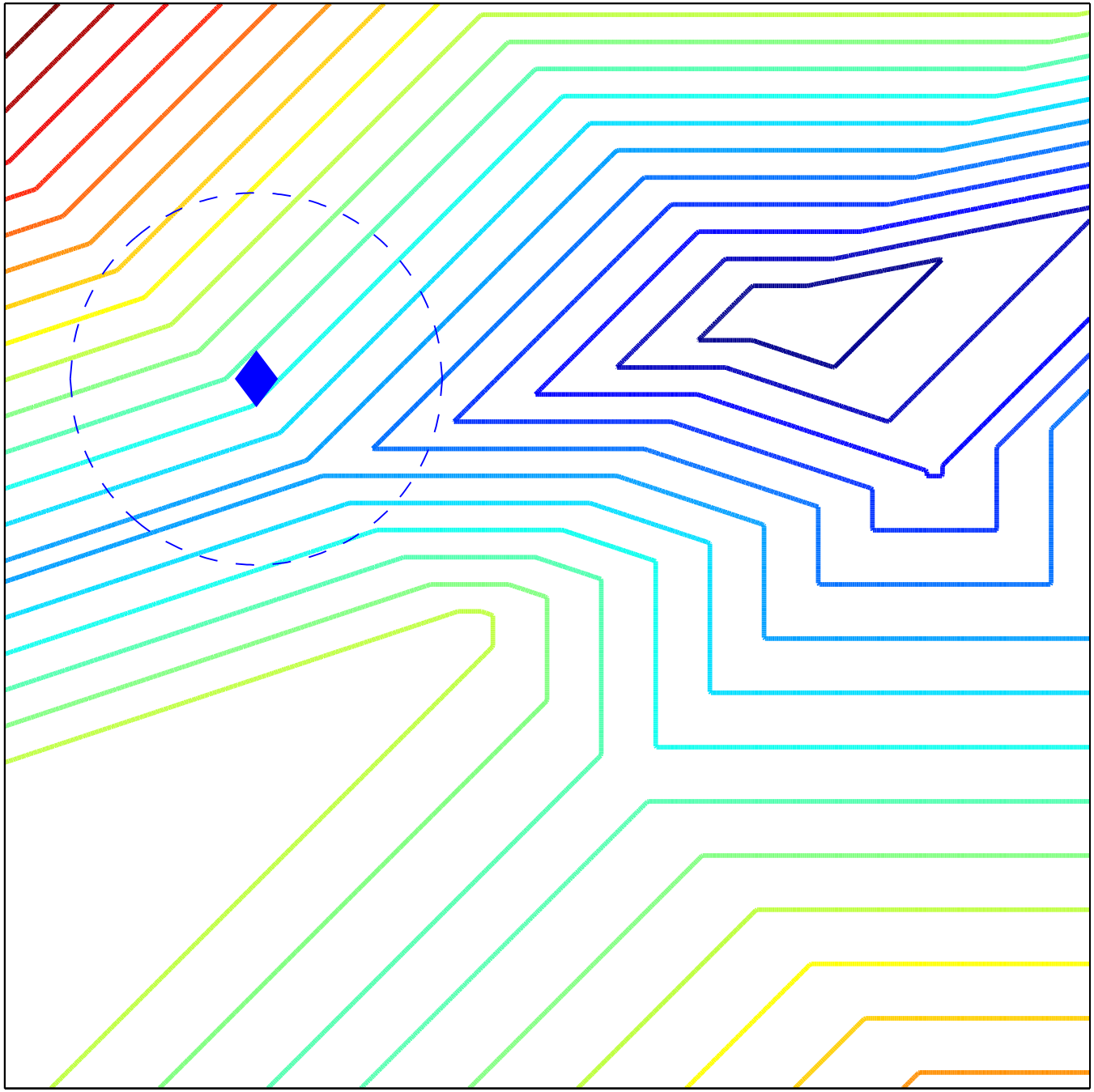


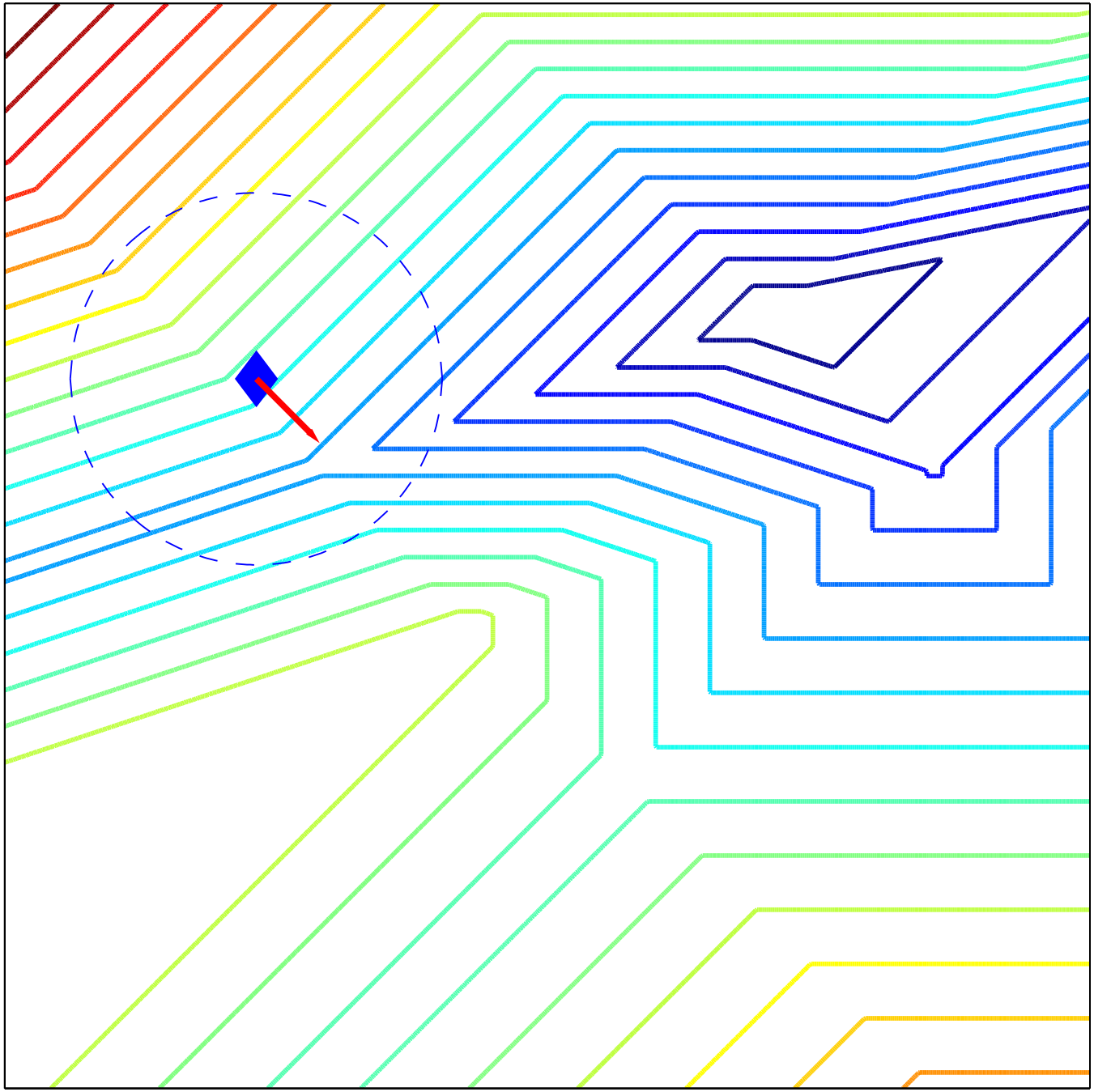


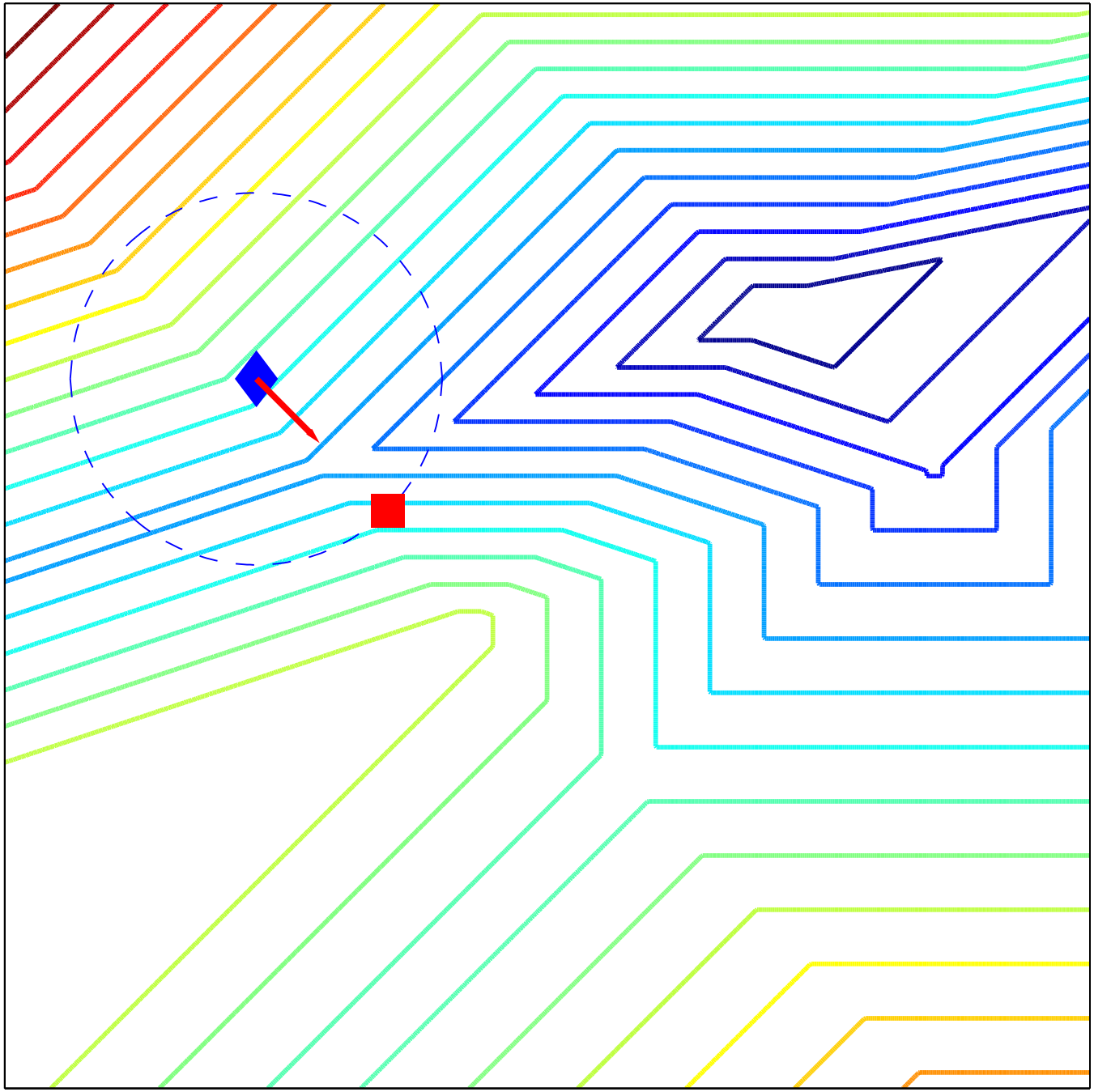


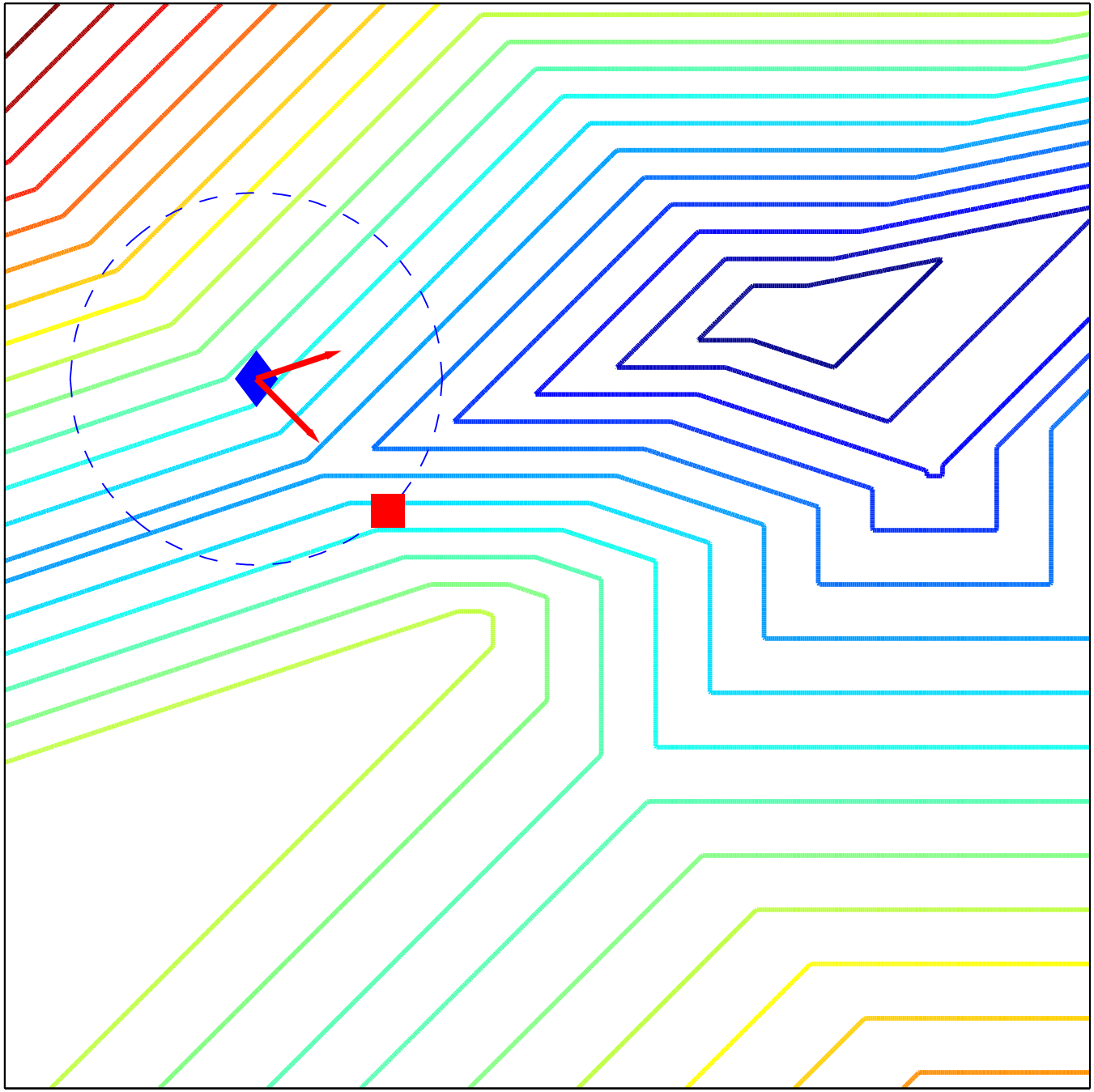


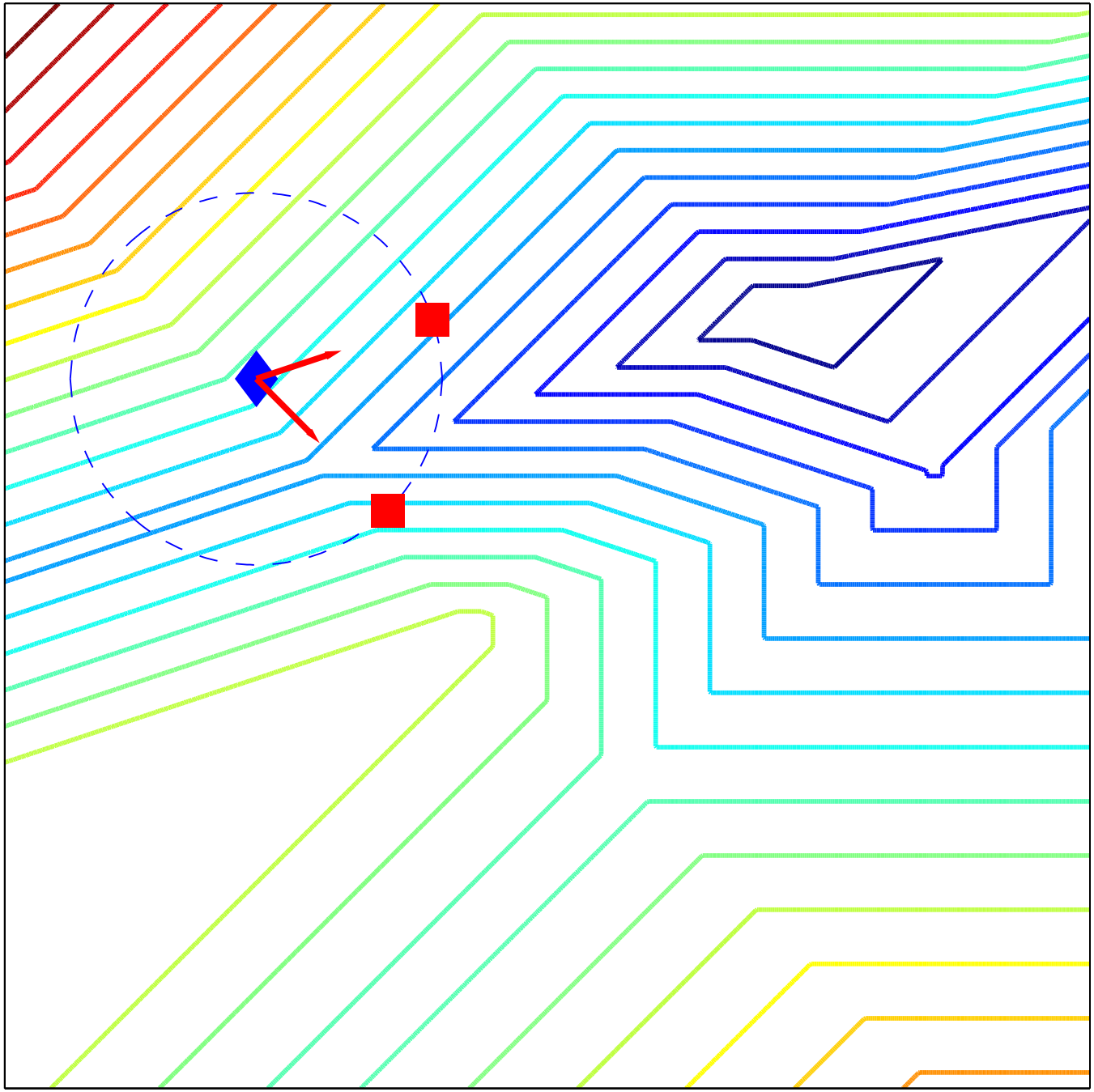












Algorithm components

- Generator set $\mathfrak{G}^k$

Algorithm components

- Generator set $\mathfrak{G}^k$

- Smooth master model m_k^f

Algorithm components

- Generator set $\mathfrak{G}^k$
- Smooth master model m_k^f
- Trust-region subproblem solution s^k

Algorithm components

- Generator set $\mathfrak{G}^k$
- Smooth master model m_k^f
- Trust-region subproblem solution s^k
- Measuring descent with ρ_k

Generator set

At some iterate x^k,

$$\mathfrak{G}^k \triangleq \bigcup_{i \in I_h(F(x^k))} \left\{ \nabla \psi(x^k) + \nabla M(x^k) a_i \right\}$$

where $I_h(F(x^k))$ is the set of essentially active indices of h at $F(x^k)$.

Generator set

At some iterate x^k,

$$\mathfrak{G}^k \triangleq \bigcup_{i \in I_h(F(x^k))} \left\{ \nabla \psi(x^k) + \nabla M(x^k) a_i \right\} \rightarrow \text{MS4PL-1}$$

where $I_h(F(x^k))$ is the set of essentially active indices of h at $F(x^k)$.

Generator set

At some iterate x^k,

$$\mathfrak{G}^k \triangleq \bigcup_{i \in I_h(F(x^k))} \left\{ \nabla \psi(x^k) + \nabla M(x^k) a_i \right\} \rightarrow \text{MS4PL-1}$$

where $I_h(F(x^k))$ is the set of essentially active indices of h at $F(x^k)$.

Or, given a set of points $Y = \left\{ x^k, y^2, \ldots, y^p \right\} \subset \mathcal{B}(x^k, \Delta_k)$,

$$\mathfrak{G}^k \triangleq \bigcup_{y \in Y} \bigcup_{i \in I_h(F(y))} \left\{ \nabla \psi(x^k) + \nabla M(x^k) a_i \right\}$$

Generator set

At some iterate x^k,

$$\mathfrak{G}^k \triangleq \bigcup_{i \in I_h(F(x^k))} \left\{ \nabla \psi(x^k) + \nabla M(x^k) a_i \right\} \rightarrow \text{MS4PL-1}$$

where $I_h(F(x^k))$ is the set of essentially active indices of h at $F(x^k)$.

Or, given a set of points $Y = \left\{ x^k, y^2, \ldots, y^p \right\} \subset \mathcal{B}(x^k, \Delta_k)$,

$$\mathfrak{G}^k \triangleq \bigcup_{y \in Y} \bigcup_{i \in I_h(F(y))} \left\{ \nabla \psi(x^k) + \nabla M(x^k) a_i \right\} \rightarrow \text{MS4PL-2}$$

Generator set

At some iterate x^k,

$$\mathfrak{G}^k \triangleq \bigcup_{i \in I_h(F(x^k))} \left\{ \nabla \psi(x^k) + \nabla M(x^k) a_i \right\} \rightarrow \text{MS4PL-1}$$

where $I_h(F(x^k))$ is the set of essentially active indices of h at $F(x^k)$.

Or, given a set of points $Y = \left\{ x^k, y^2, \ldots, y^p \right\} \subset \mathcal{B}(x^k, \Delta_k)$,

$$\mathfrak{G}^k \triangleq \bigcup_{y \in Y} \bigcup_{i \in I_h(F(y))} \left\{ \nabla \psi(x^k) + \nabla M(x^k) a_i \right\} \rightarrow \text{MS4PL-2}$$

Assumption

The set $\mathfrak{G}^k$ *satisfies* MS4PL-1 $\subseteq \mathfrak{G}^k \subseteq$ MS4PL-2.

Smooth master model

Our model gradients around iterate x^k satisfy

$$g^k \triangleq \mathbf{proj}\left(0, \mathbf{co}\left(\mathfrak{G}^k\right)\right) \in \mathbf{co}\left(\mathfrak{G}^k\right),$$

Let λ^* be the corresponding coefficients so that $g^k = G^k \lambda^*$.

Smooth master model

Our model gradients around iterate x^k satisfy

$$g^k \triangleq \mathbf{proj}\left(0, \mathbf{co}\left(\mathfrak{G}^k\right)\right) \in \mathbf{co}\left(\mathfrak{G}^k\right),$$

Let λ^* be the corresponding coefficients so that $g^k = G^k\lambda^*$.

Define

$$A^k \triangleq \begin{bmatrix} | & & | \\ a_{j_1} & \cdots & a_{j_t} \\ | & & | \end{bmatrix},$$

and set $w^k = A^k\lambda^*$. Define the smooth ***master model*** $m_k^f : \mathbb{R}^n \to \mathbb{R}$,

$$m_k^f(x) \triangleq \psi(x) + \sum_{i=1}^{p} w_i^k m^{F_i}(x) + \sum_{i=1}^{p} \lambda_i^* b_{j_i}.$$

Trust region subproblem

Approximately solve

$$\begin{aligned} &\underset{s}{\text{minimize}}\ m_k^f(x^k + s) \\ &\text{subject to: } s \in \mathcal{B}(0, \Delta_k) \end{aligned}$$

to obtain a solution s satisfying

$$\psi(x^k) - \psi(x^k + s) + \left\langle M(x^k) - M(x^k + s), w^k \right\rangle \geq \frac{\kappa_{\mathrm{d}}}{2} \|g^k\| \min\left\{ \Delta_k, \frac{\|g^k\|}{\kappa_{\mathrm{mh}}} \right\}.$$

Measuring descent

- Descent is measured using some selection function $h^{(k)}$ and not h

Measuring descent

- Descent is measured using some selection function $h^{(k)}$ and not h

- Must ensure information about $h^{(k)}$ is in $\mathfrak{G}^k$ before taking a step

Measuring descent

- Descent is measured using some selection function $h^{(k)}$ and not h
- Must ensure information about $h^{(k)}$ is in $\mathfrak{G}^k$ before taking a step
- $h^{(k)}$ must satisfy

$$h^{(k)}(F(x^k)) \leq h(F(x^k)) \qquad \text{and} \qquad h^{(k)}(F(x^k + s^k)) \geq h(F(x^k + s^k)),$$

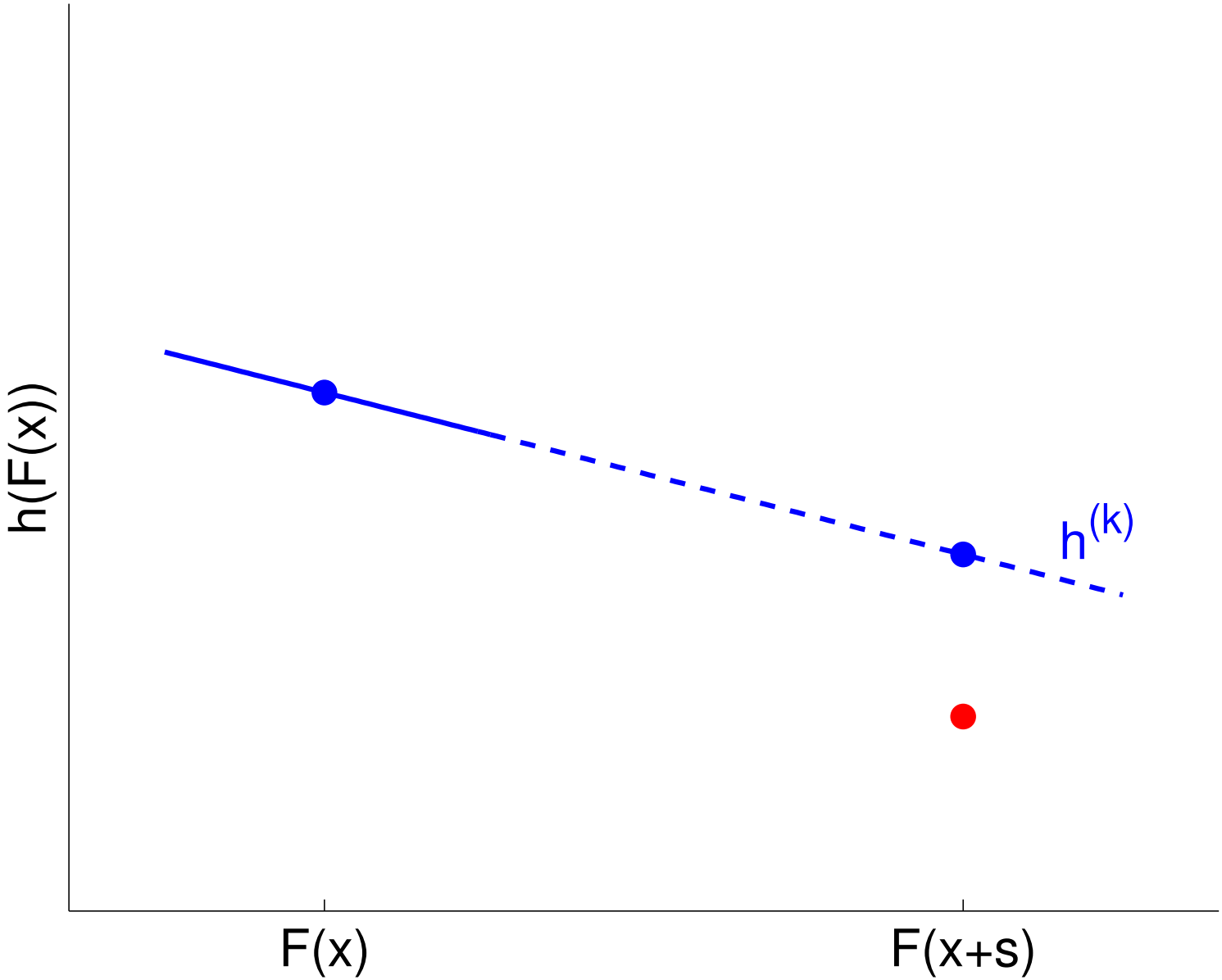
Measuring descent

- Descent is measured using some selection function $h^{(k)}$ and not h
- Must ensure information about $h^{(k)}$ is in $\mathfrak{G}^k$ before taking a step
- $h^{(k)}$ must satisfy

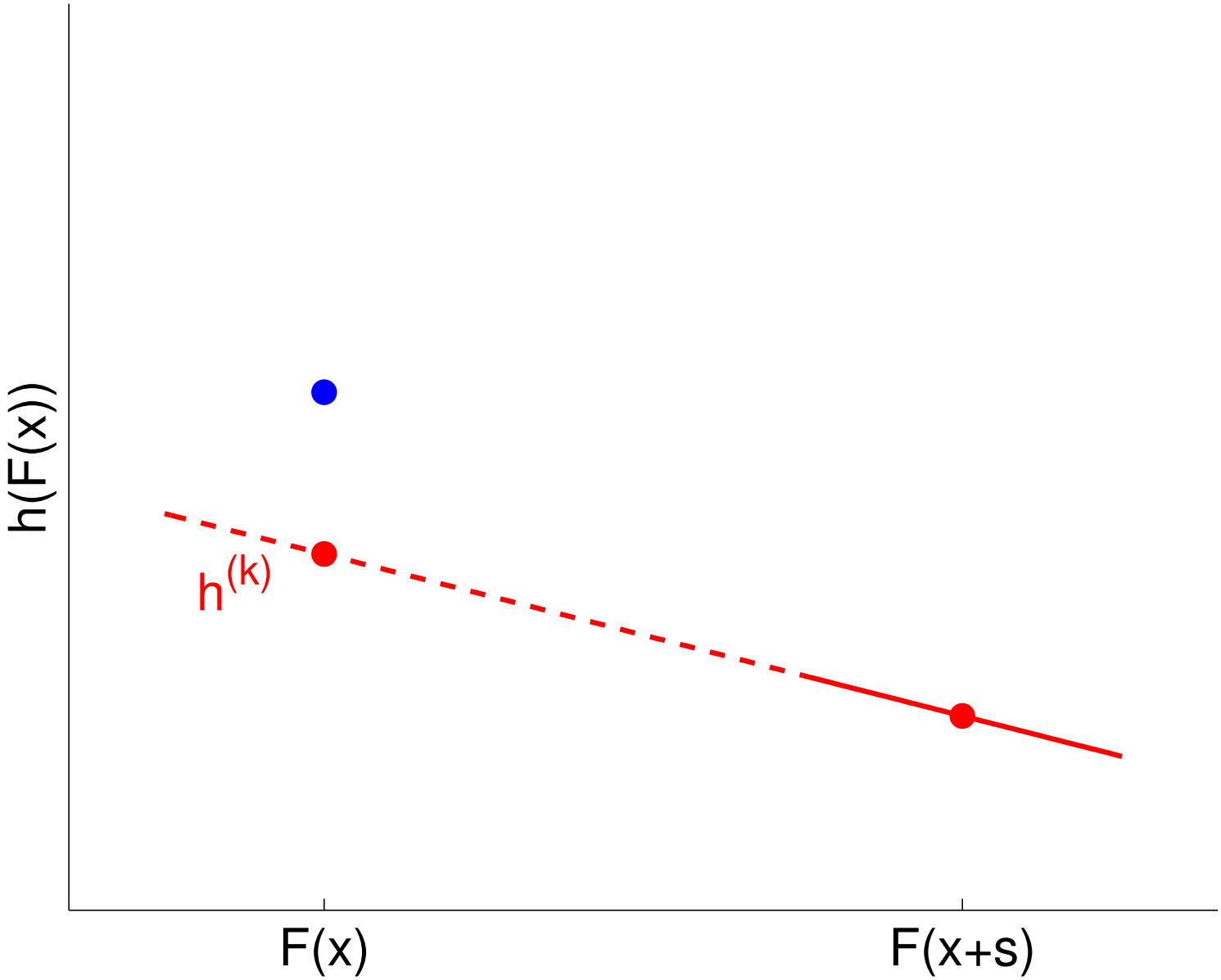
$$h^{(k)}(F(x^k)) \leq h(F(x^k)) \qquad \text{and} \qquad h^{(k)}(F(x^k + s^k)) \geq h(F(x^k + s^k)),$$

- $$\rho_k \triangleq \frac{\psi(x^k) - \psi(x^k + s^k) + h^{(k)}(F(x^k)) - h^{(k)}(F(x^k + s^k))}{\psi(x^k) - \psi(x^k + s^k) + \langle M(x^k) - M(x^k + s^k), a^{(k)} \rangle}$$

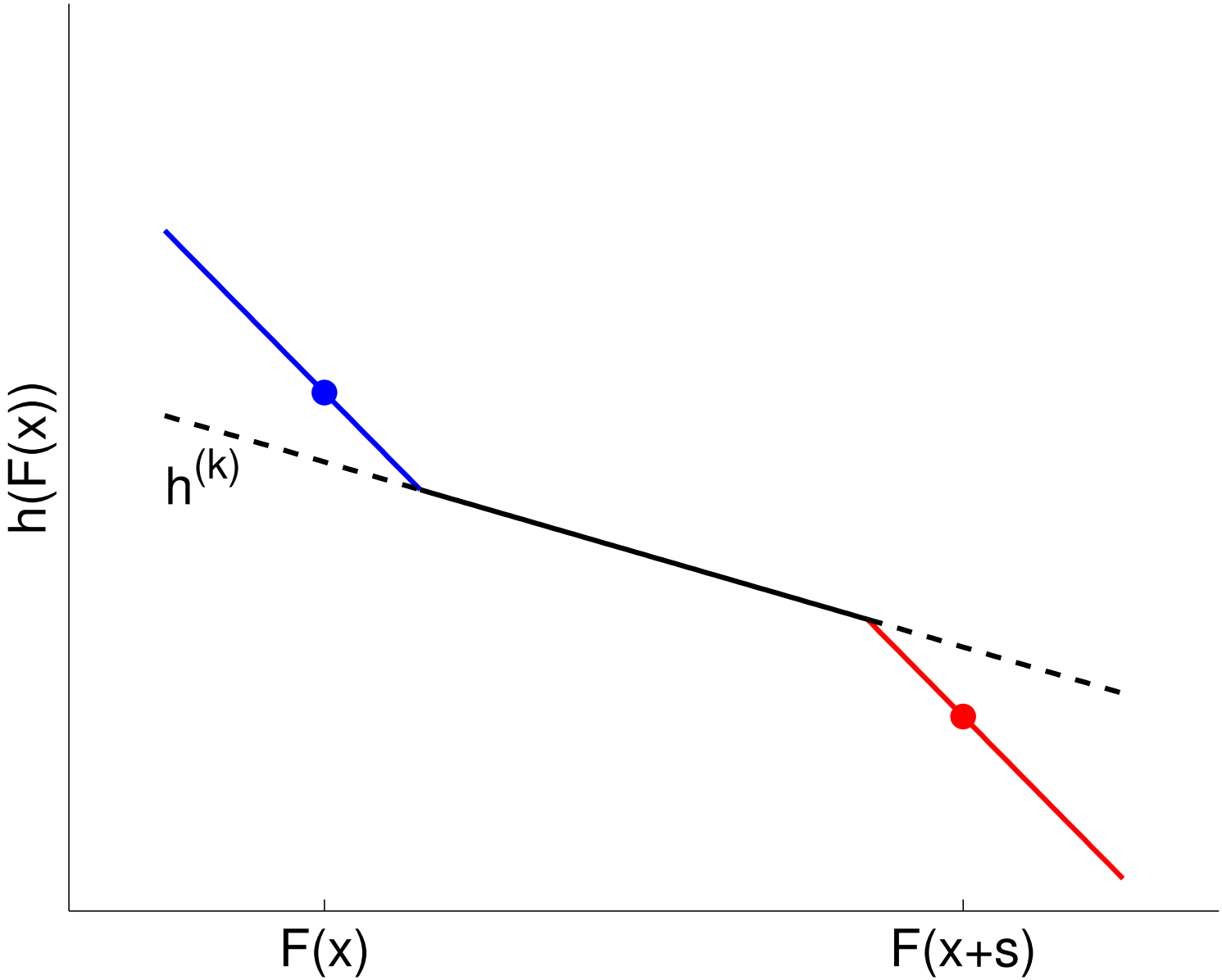
Examples of $h^{(k)}$



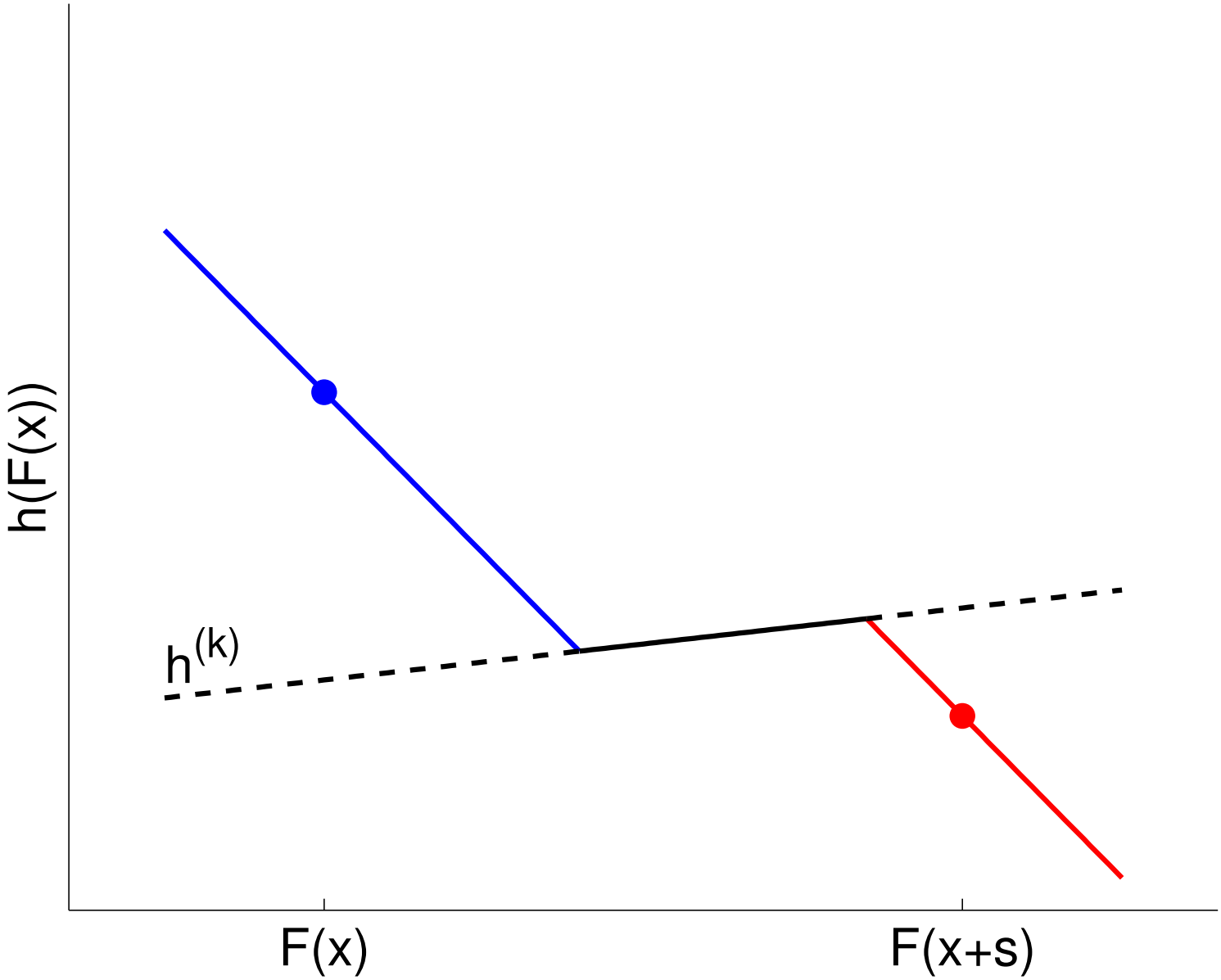
Examples of $h^{(k)}$



Examples of $h^{(k)}$



Examples of $h^{(k)}$



Algorithm components

- Generator set $\mathfrak{G}^k$
- Smooth master model m_k^f
- Trust-region subproblem solution s^k
- Measuring descent with ρ_k

Algorithm MS4PL

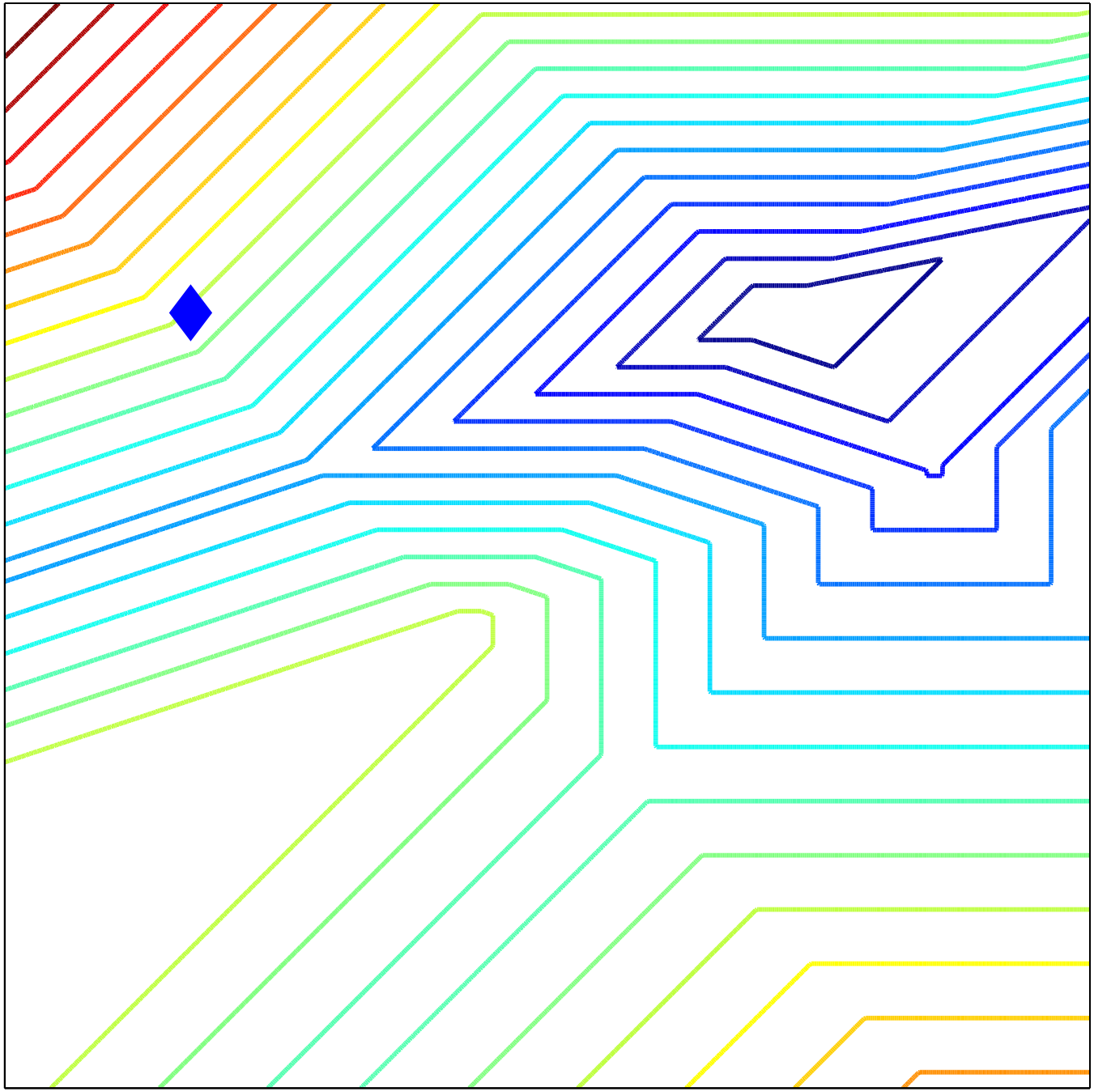
Choose x^0 **and** Δ_0
for $k = 0, 1, 2, \ldots$ **do**
  Build p **component models** m^{F_i} **fully linear on** $\mathcal{B}(x^k, \Delta_k)$
  Form $\nabla M(x^k)$ **using** $\nabla m^{F_i}(x^k)$ **and construct** $\mathfrak{G}^k \subset \mathbb{R}^n$
  $\rho_k \leftarrow -\infty$
  while $\rho_k = -\infty$ **do**
    if $\Delta_k < \eta_2 \|\nabla m^f(x^k)\|$ **then**
      Approximately solve TRSP to obtain s^k
      Evaluate $F(x^k + s^k)$ **and find** $h^{(k)}$
      if $(\nabla \psi(x^k) + \nabla M(x^k)\, a^{(k)}) \in \mathfrak{G}^k$ **then**
        Calculate ρ_k
      else
        $\mathfrak{G}^k \leftarrow \mathfrak{G}^k \cup \{\nabla \psi(x^k) + \nabla M(x^k)\, a^{(k)}\}$
        Update component models m^{F_i} **and master model** m^f
    else
      break
  if $\rho_k > \eta_1 > 0$ **then**
    $x^{k+1} \leftarrow x^k + s^k$, $\Delta_{k+1} \leftarrow \min\{\gamma_{\text{inc}} \Delta_k, \Delta_{\max}\}$
  else
    $x^{k+1} \leftarrow x^k$, $\Delta_{k+1} \leftarrow \gamma_{\text{dec}} \Delta_k$

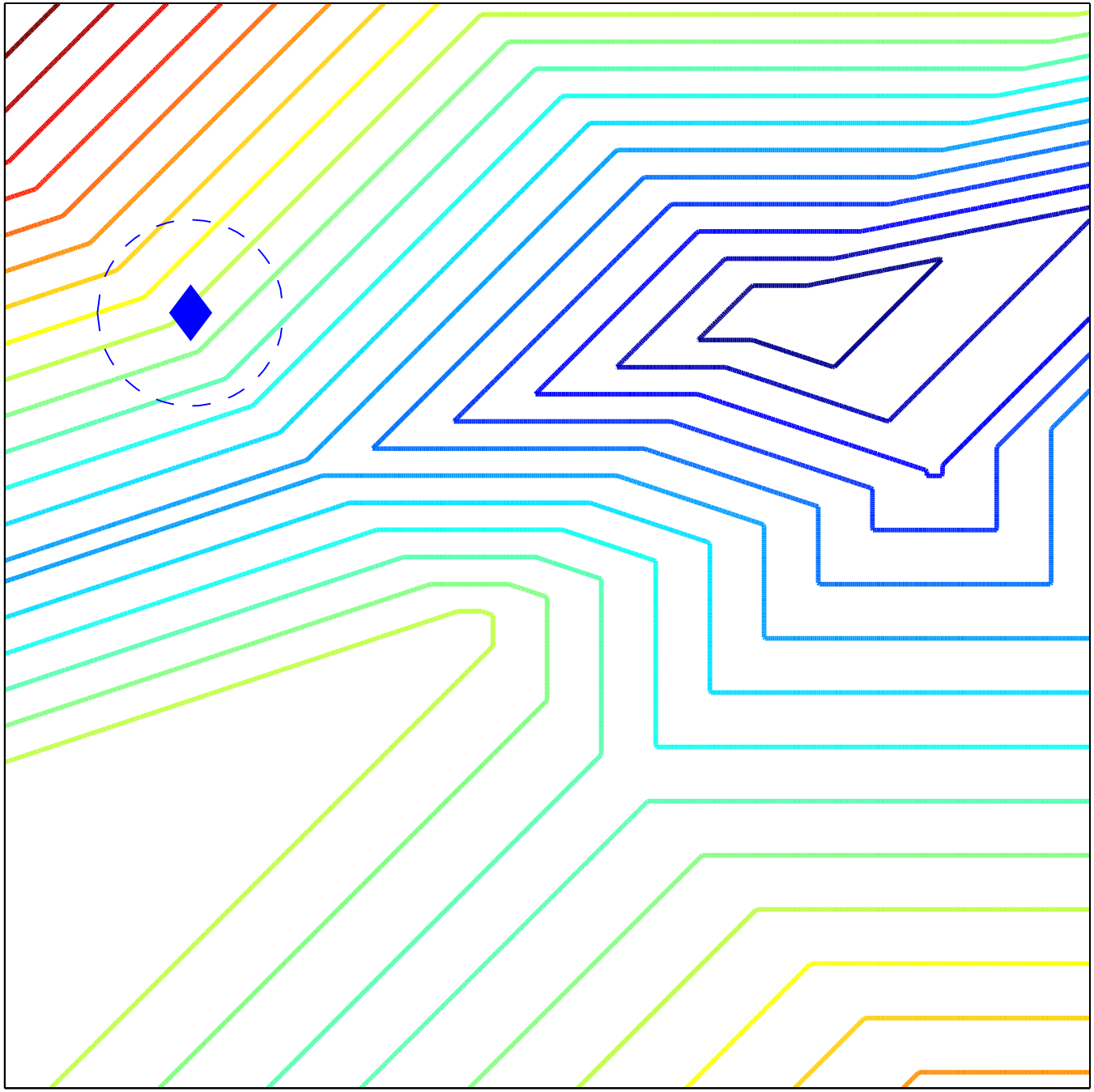
Generator set

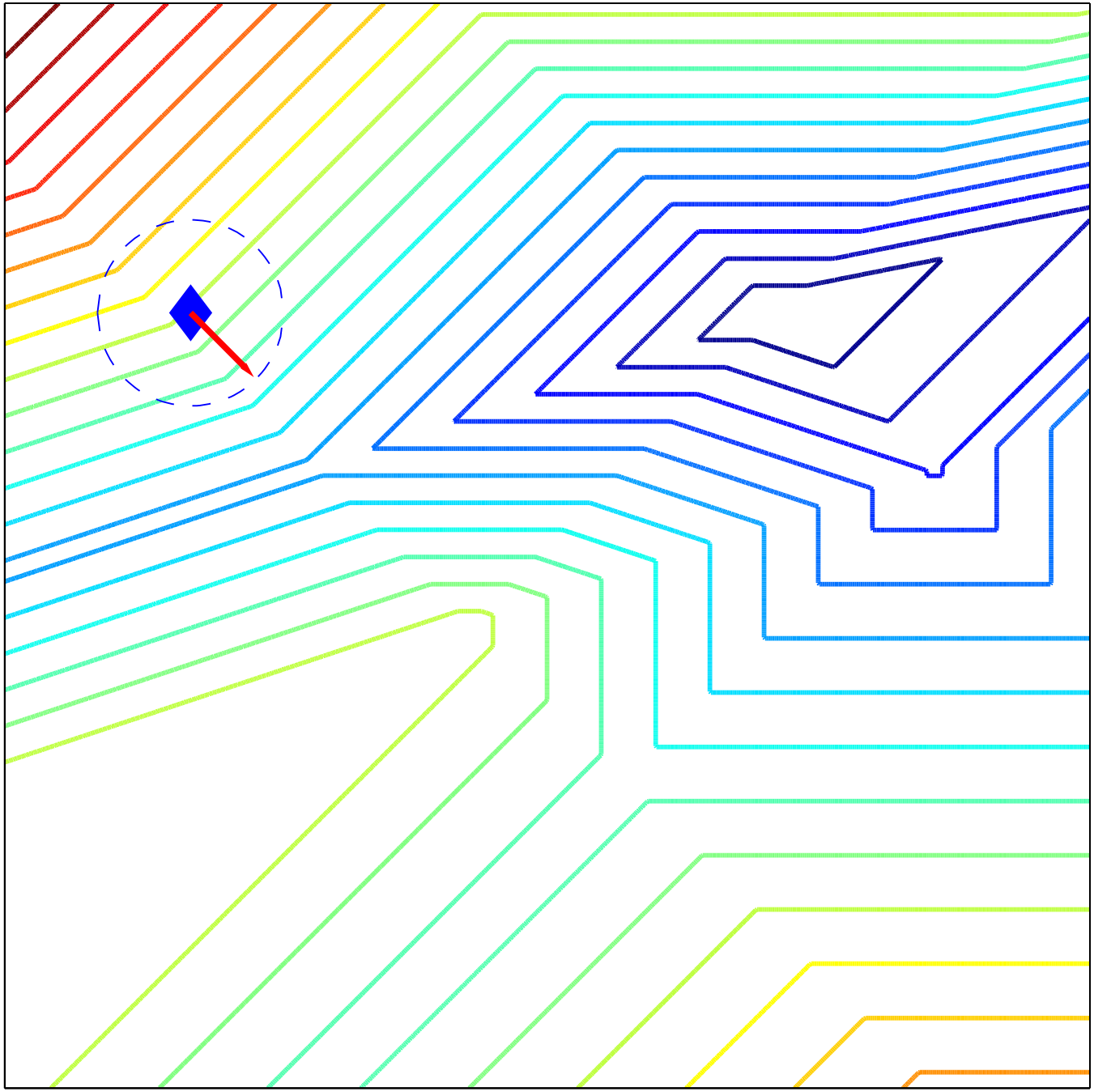
At some iterate x^k,

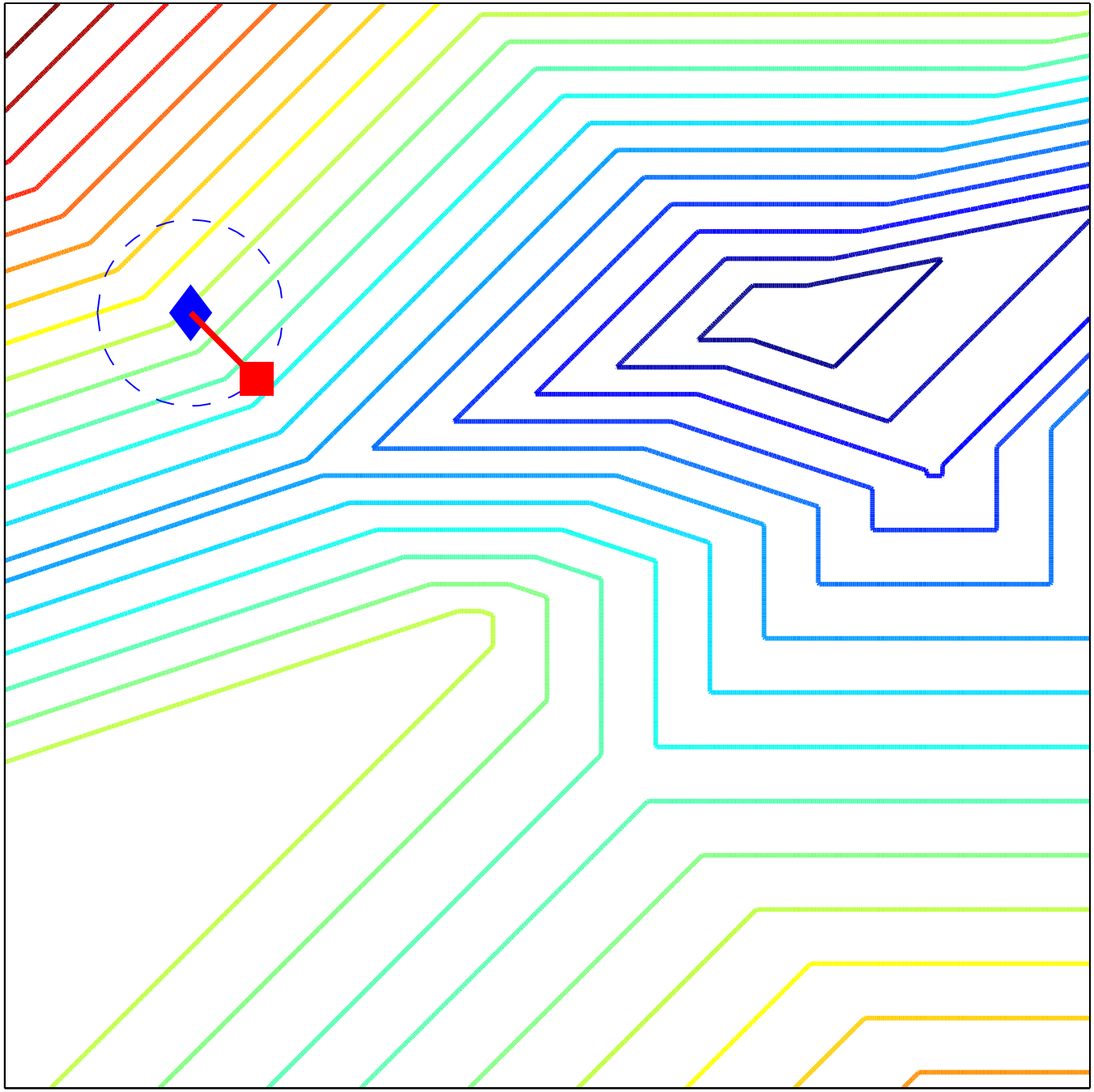
$$\mathfrak{G}^k \triangleq \bigcup_{i \in I_h(F(x^k))} \left\{ \nabla \psi(x^k) + \nabla M(x^k) a_i \right\} \rightarrow \text{MS4PL-1}$$

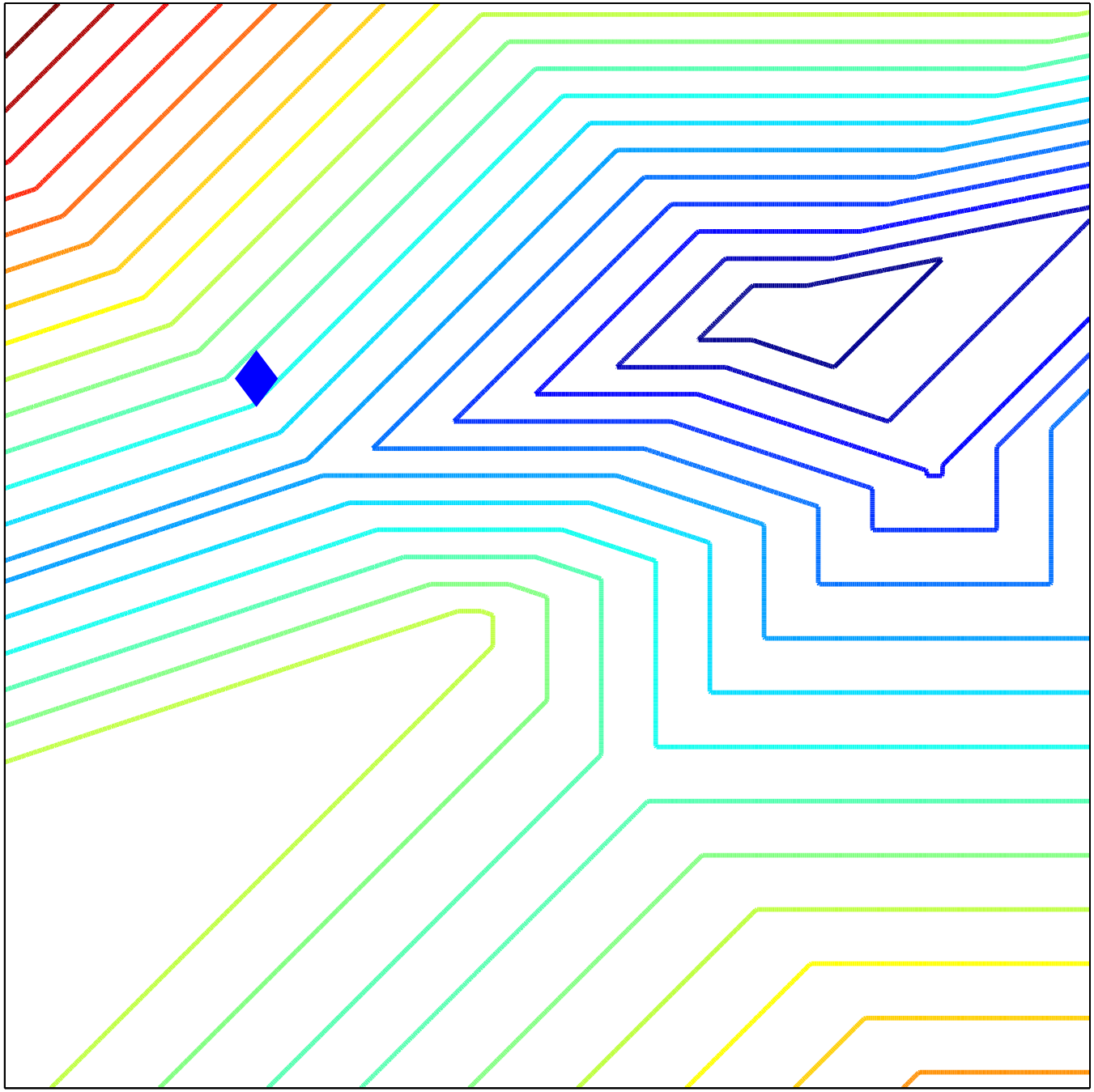
where $I_h(F(x^k))$ is the set of essentially active indices of h at $F(x^k)$.

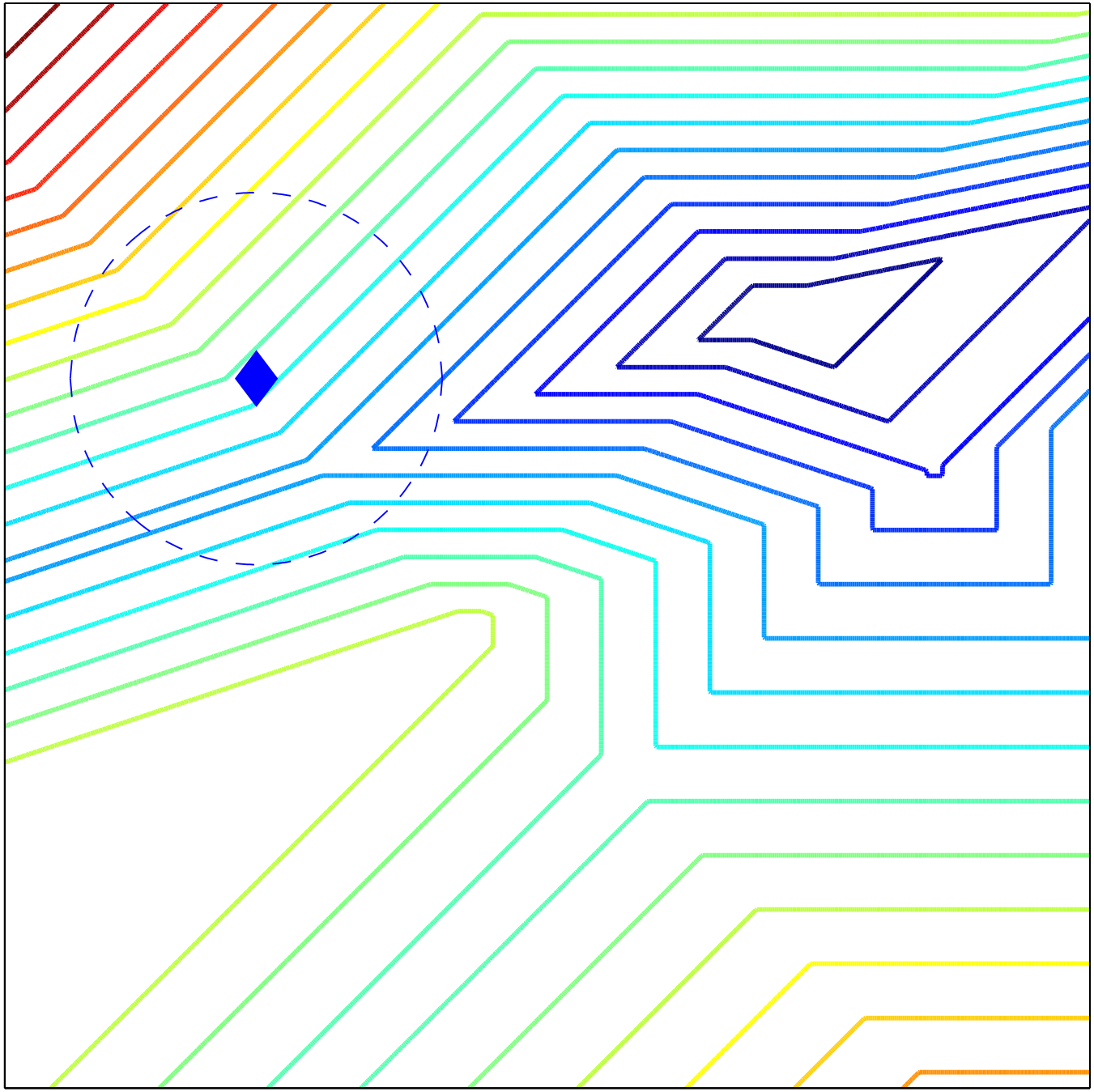


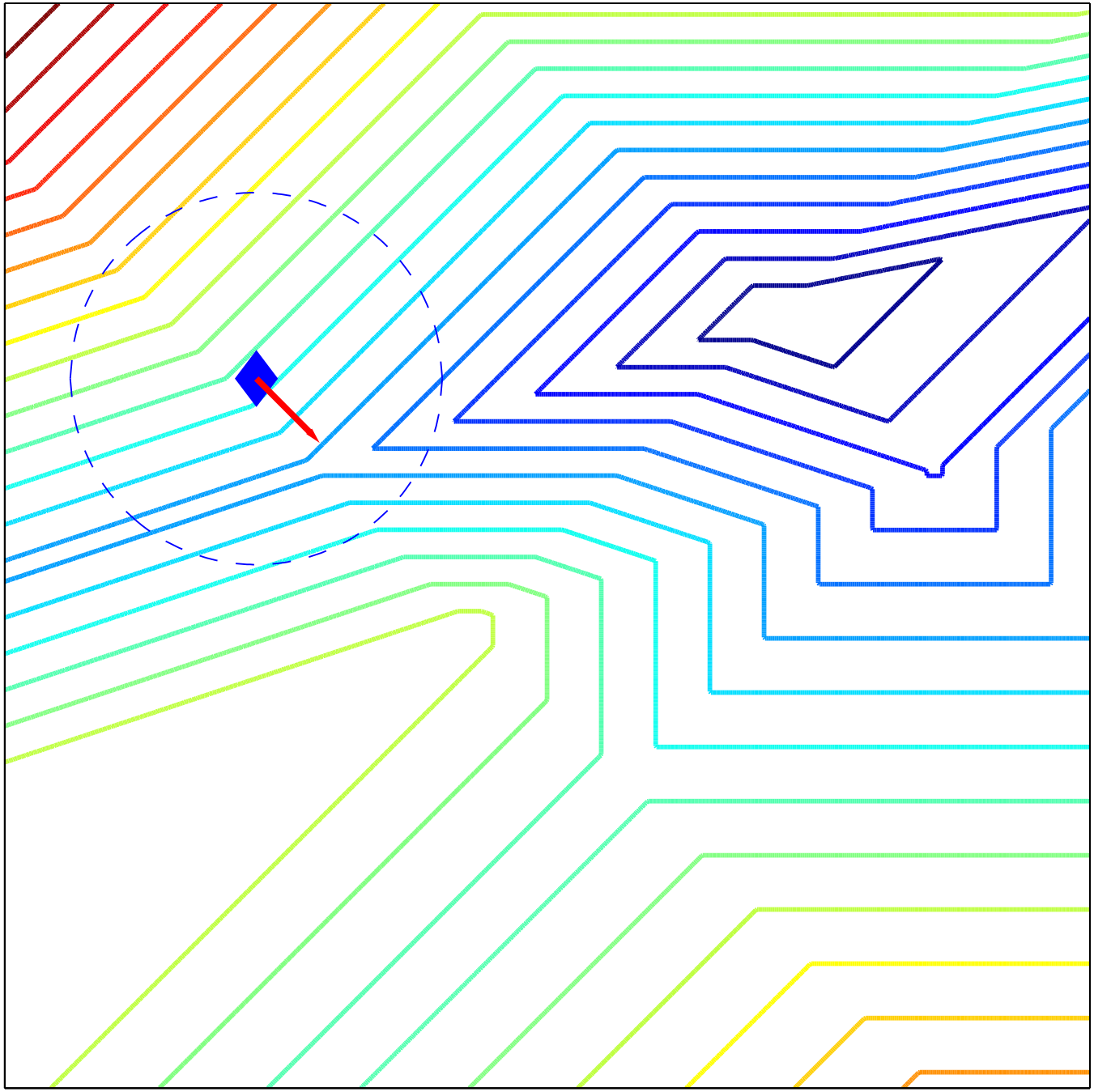


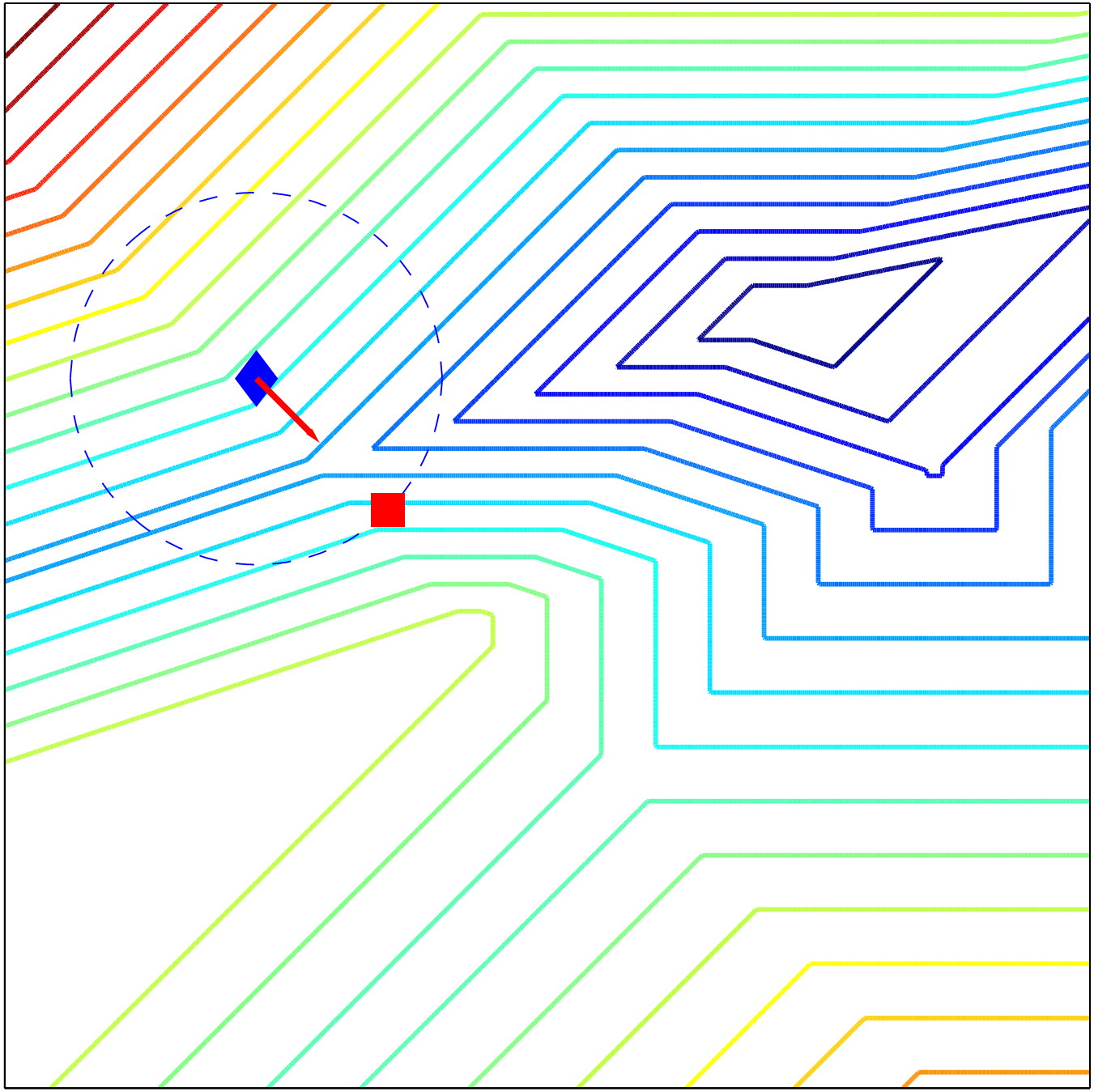


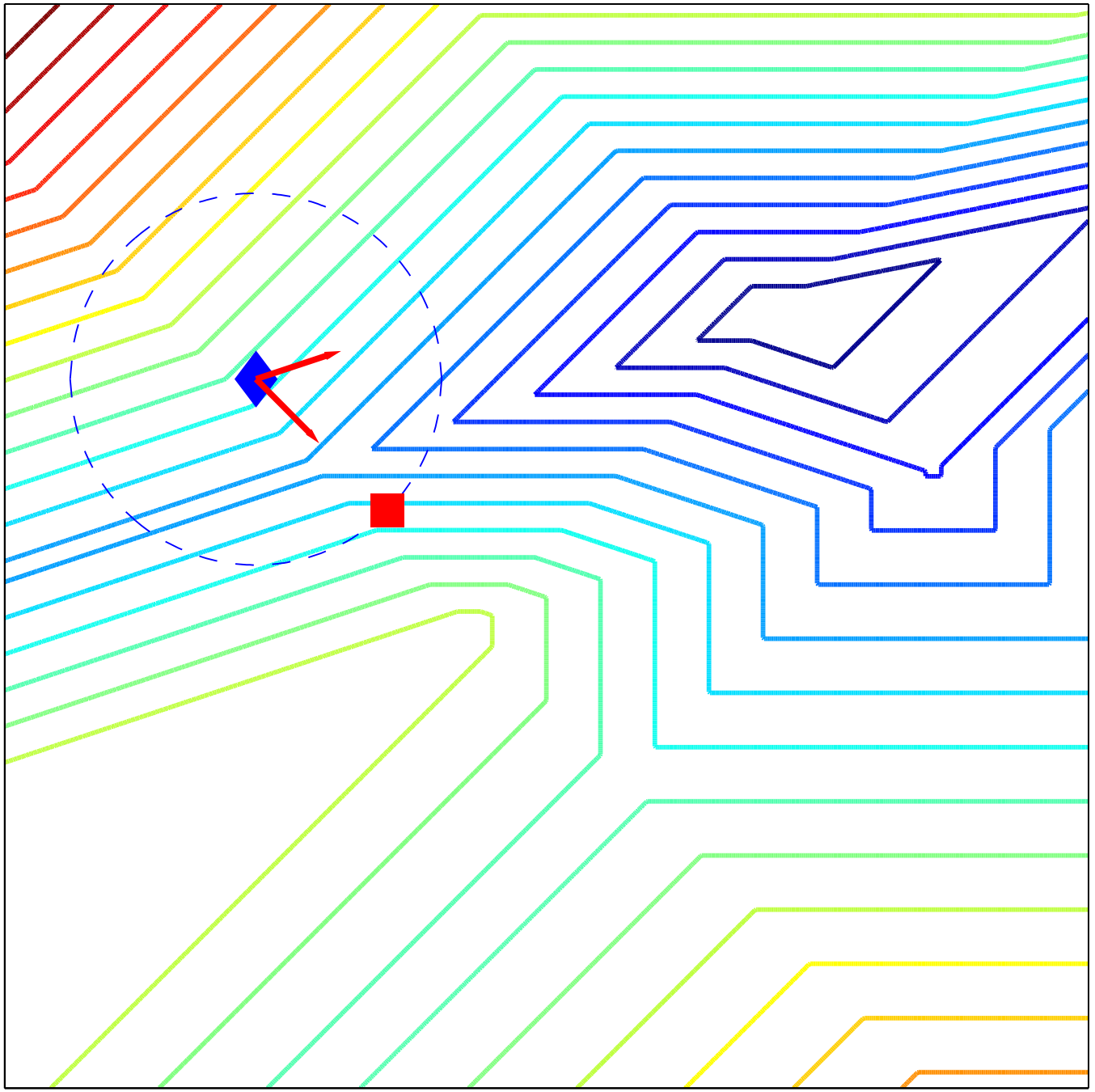


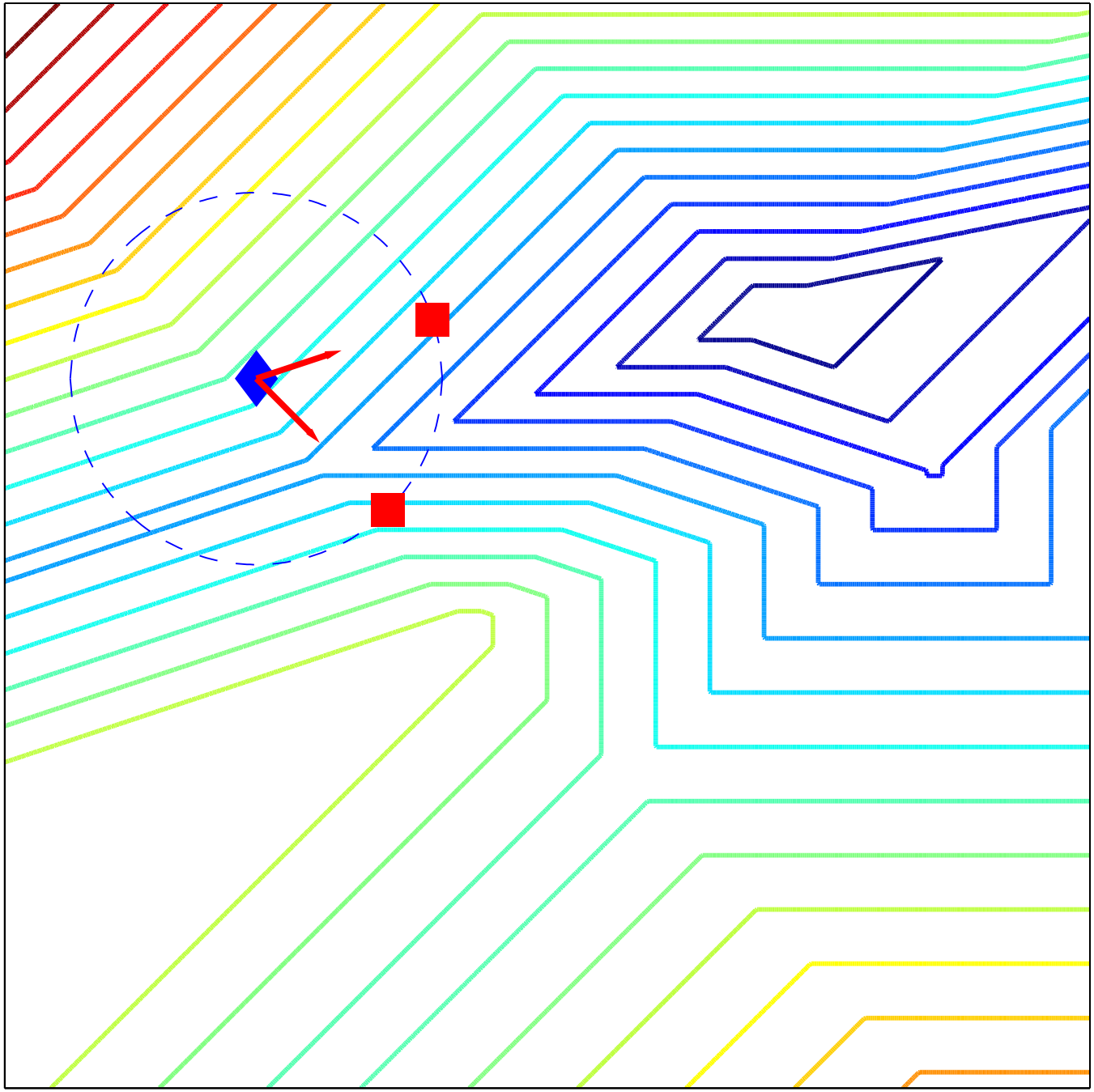


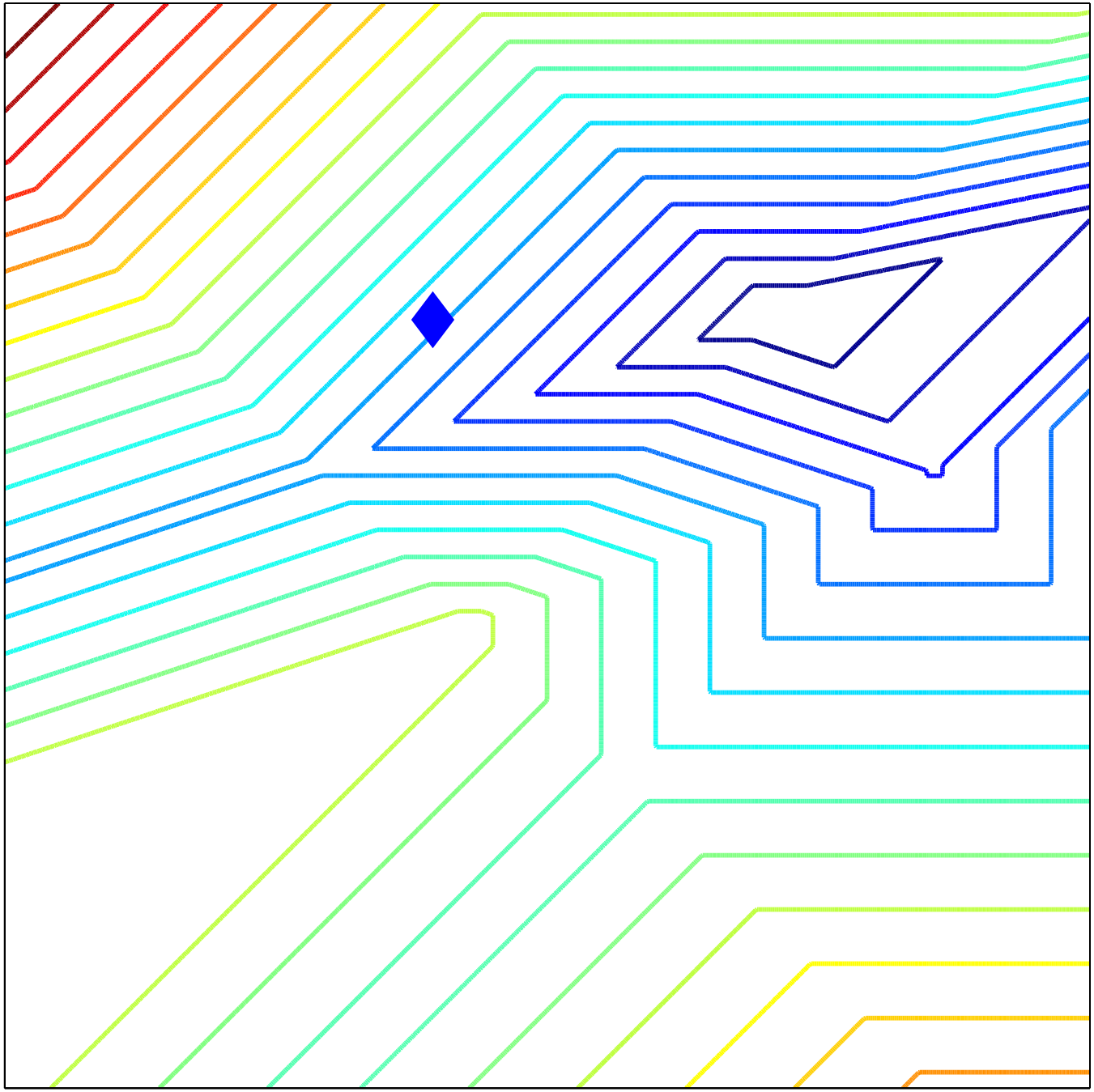


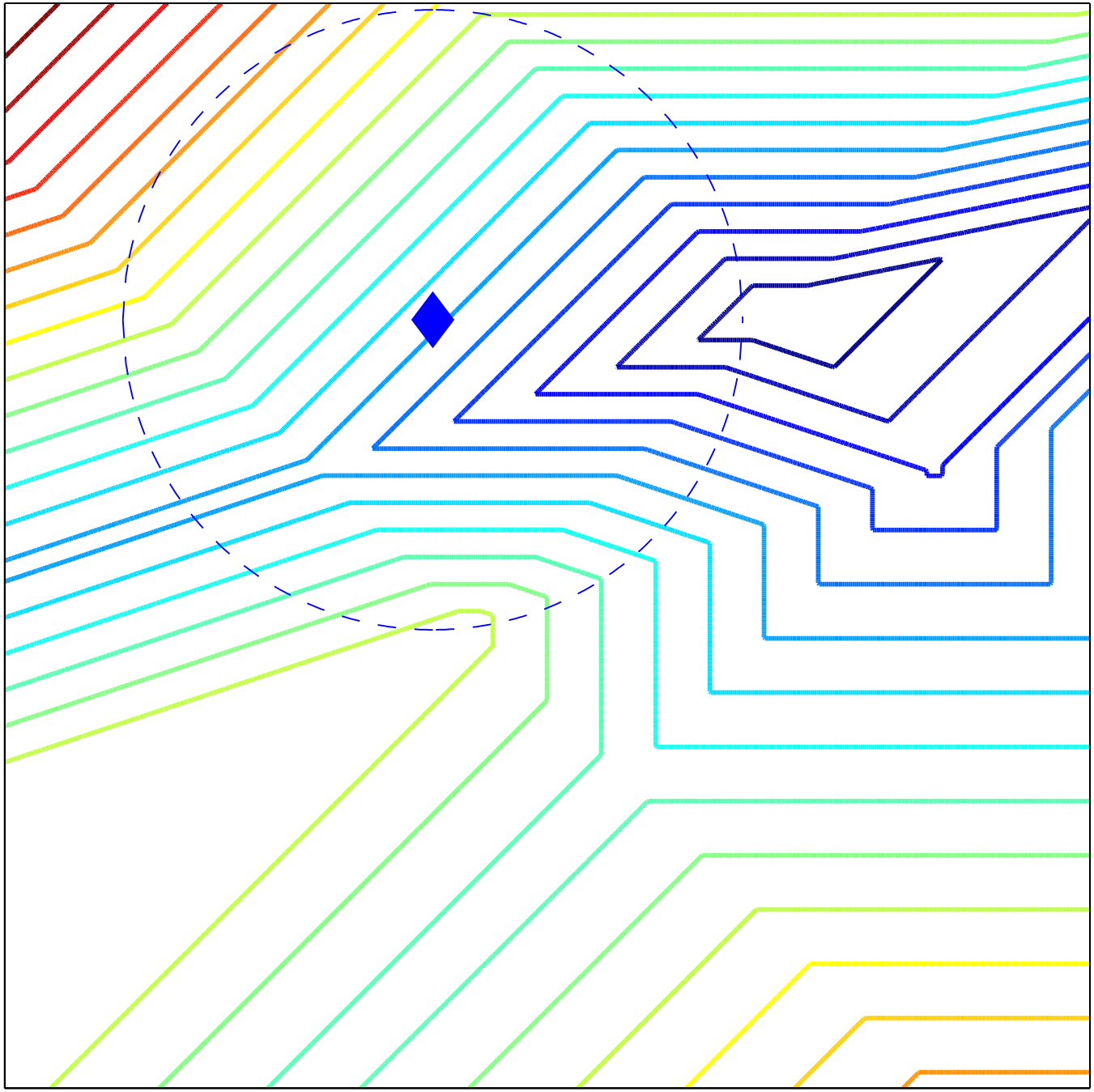


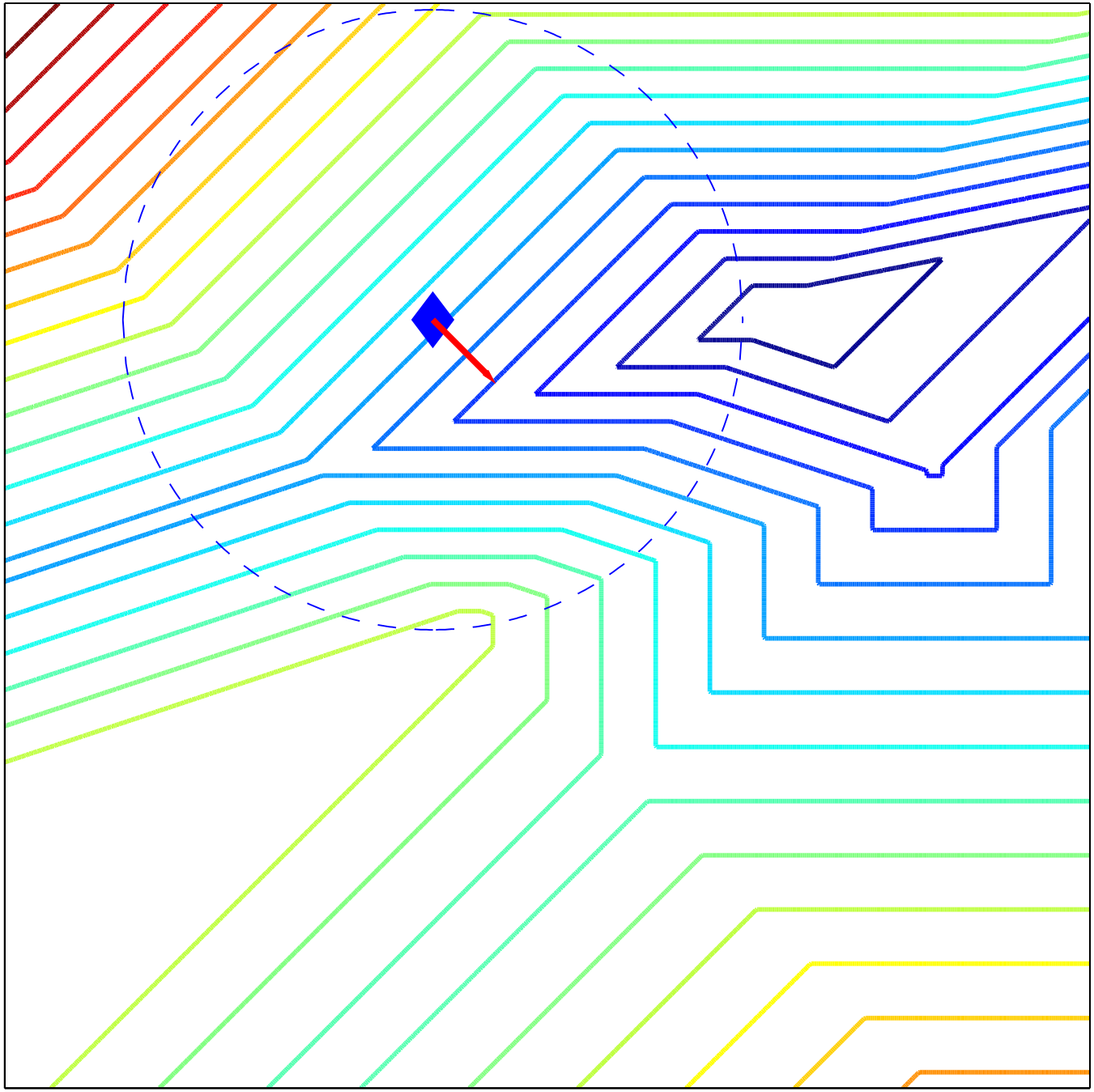


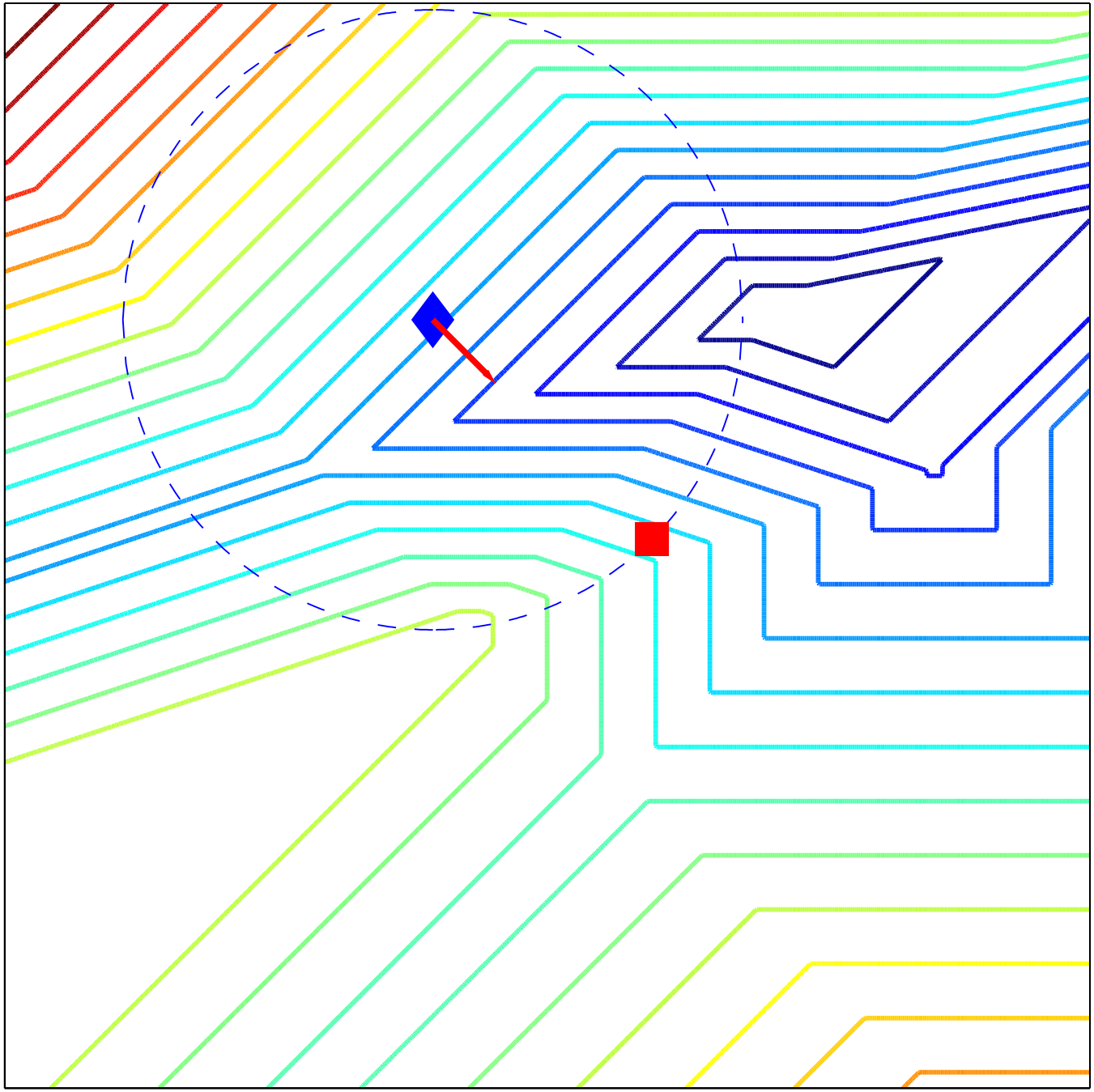


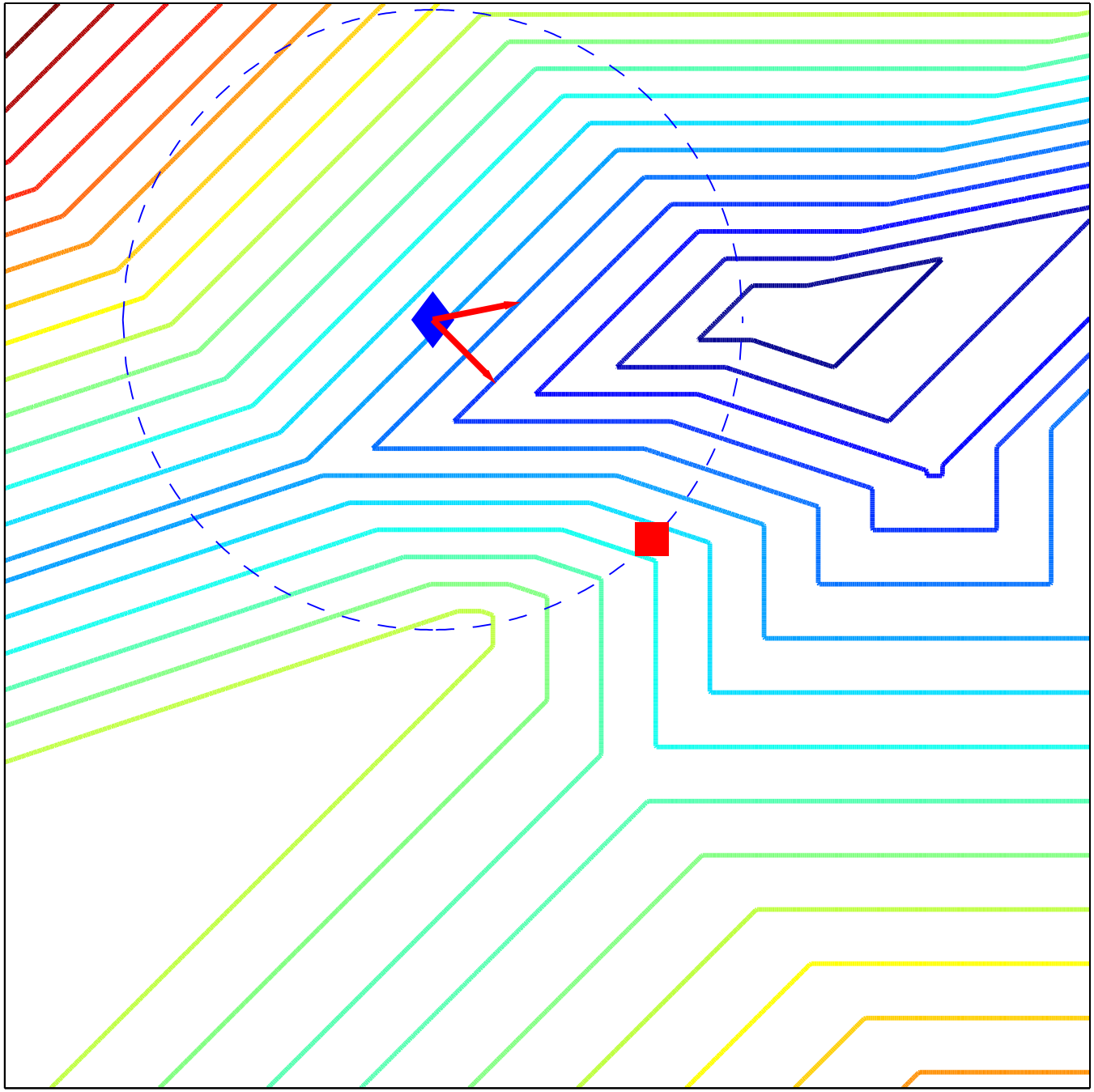


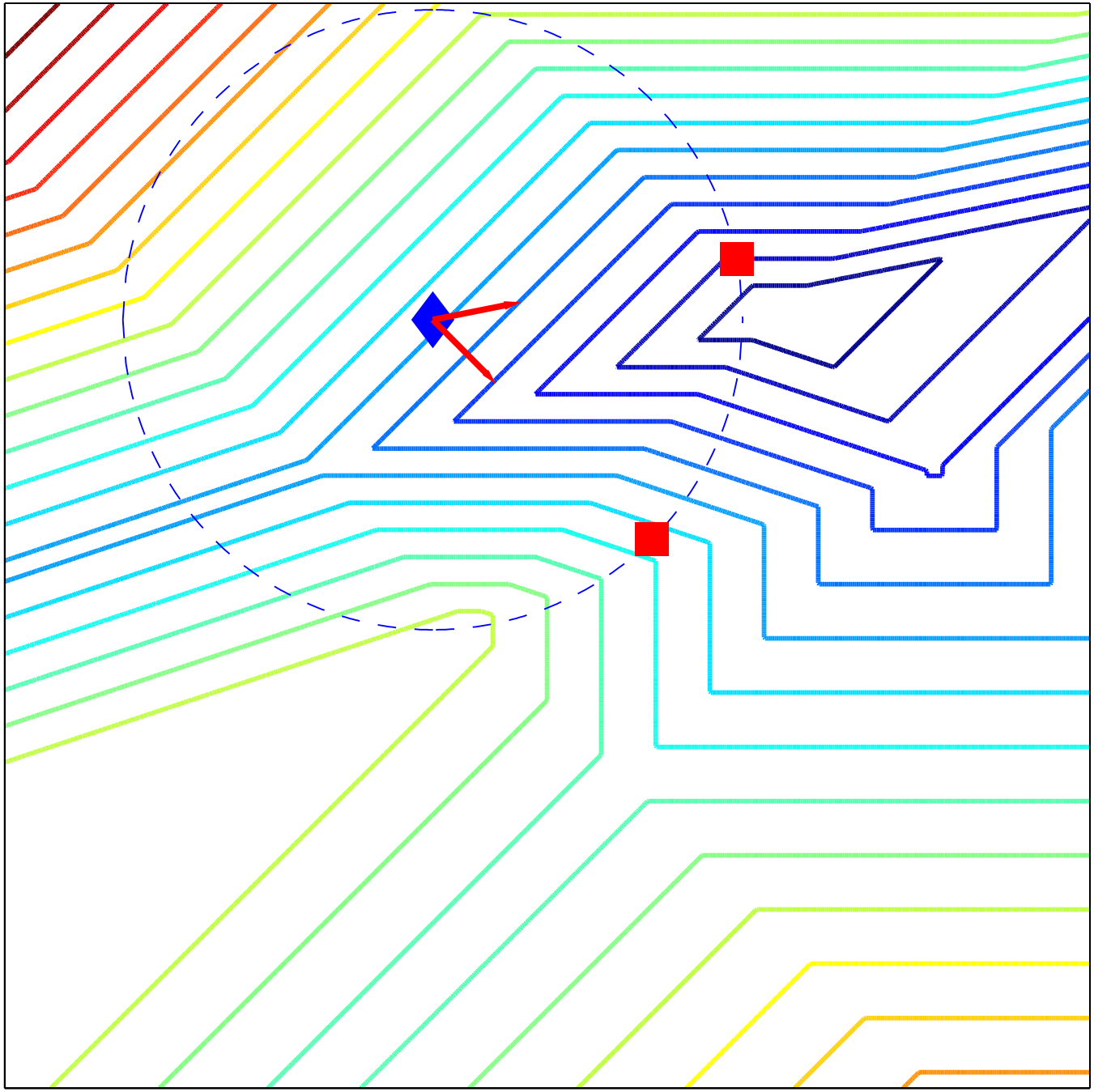


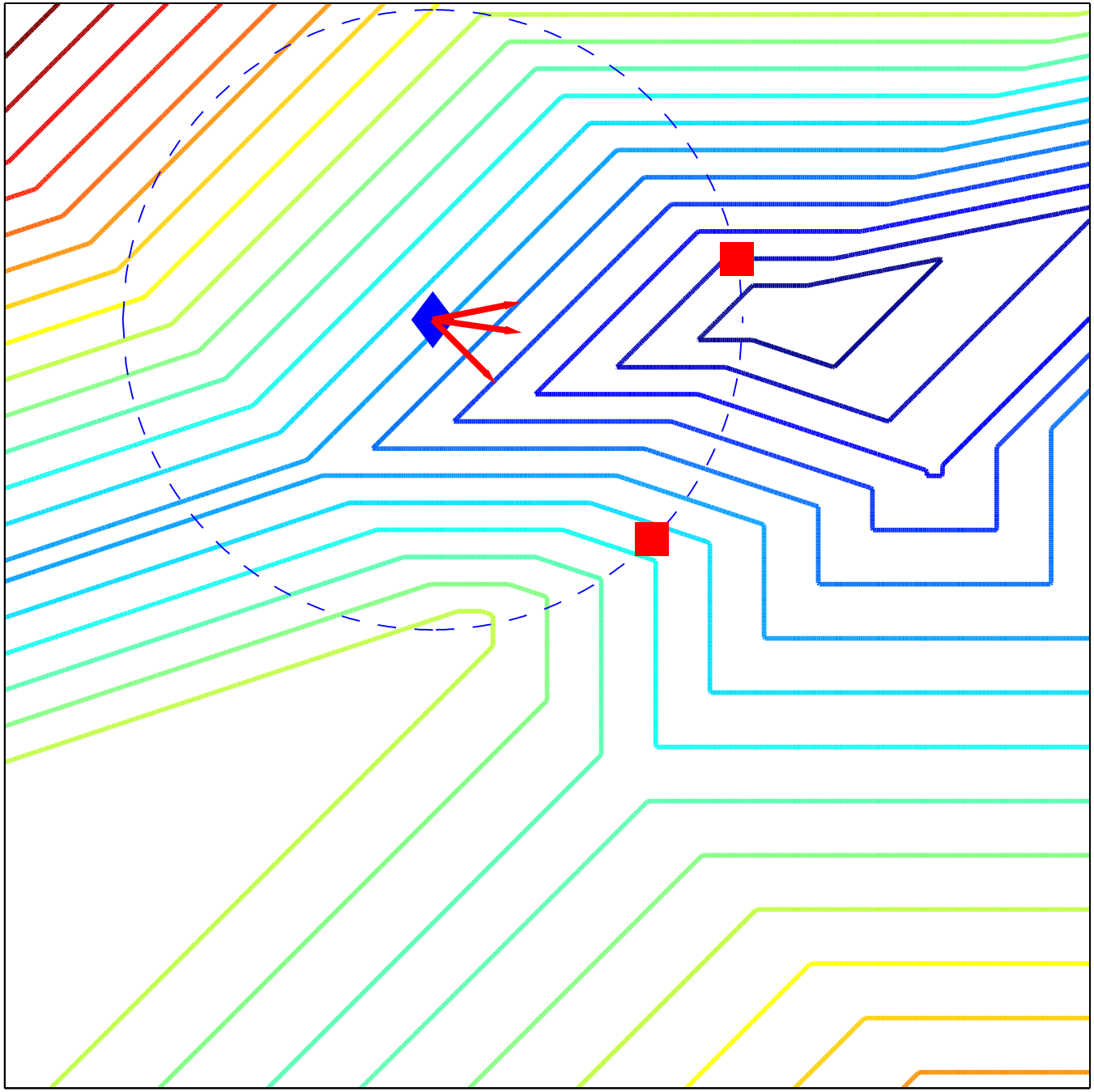


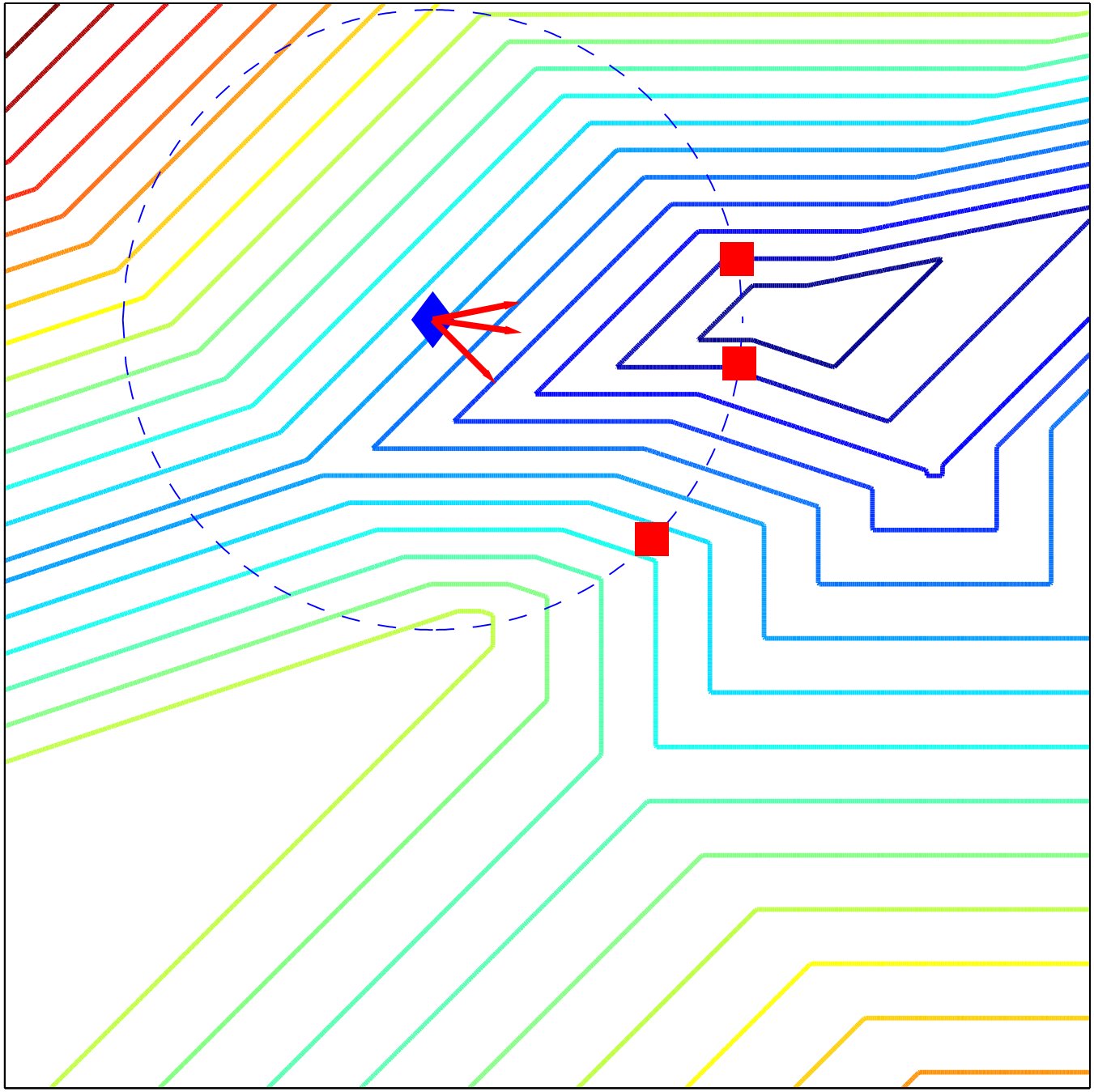


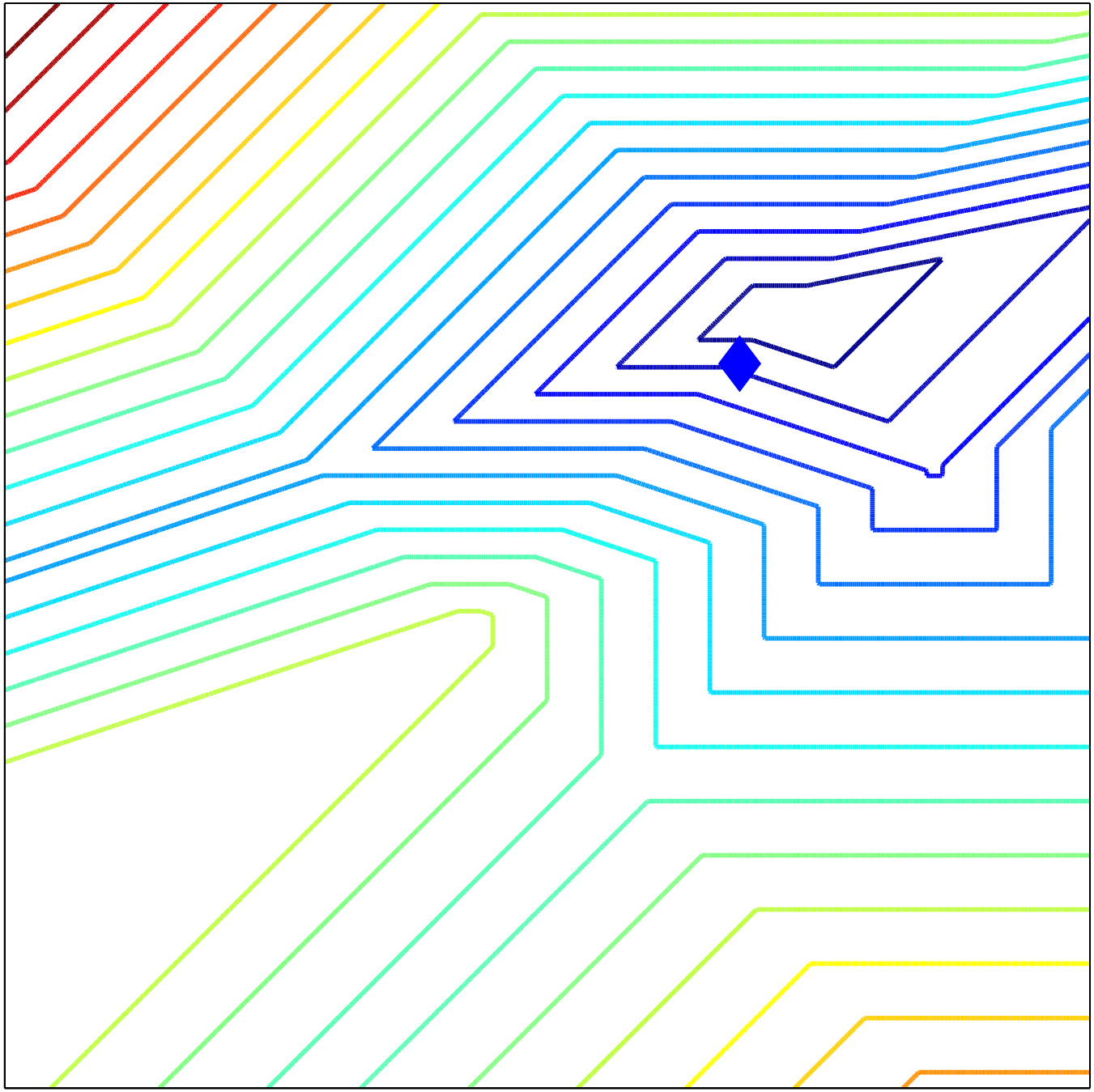


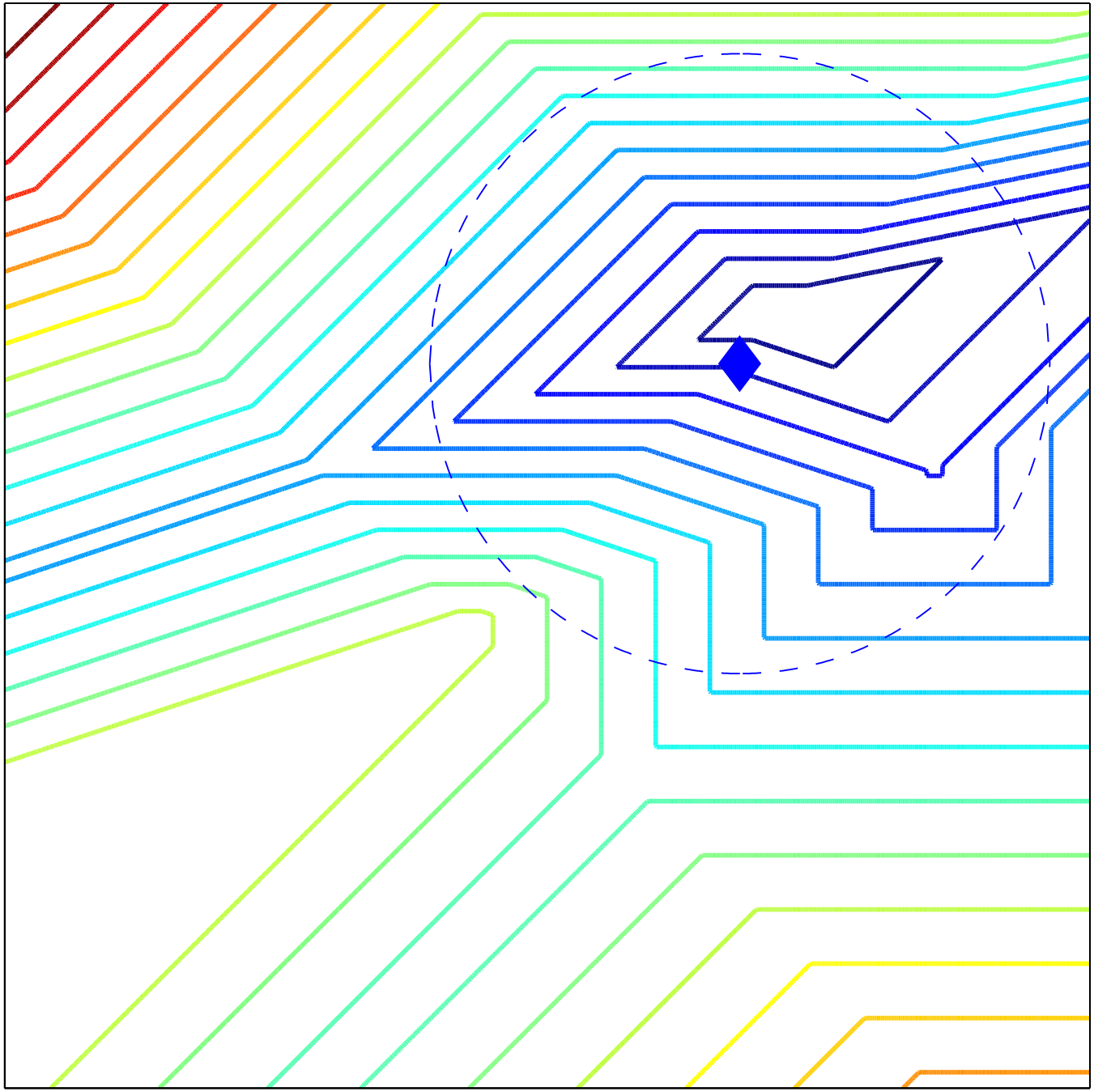


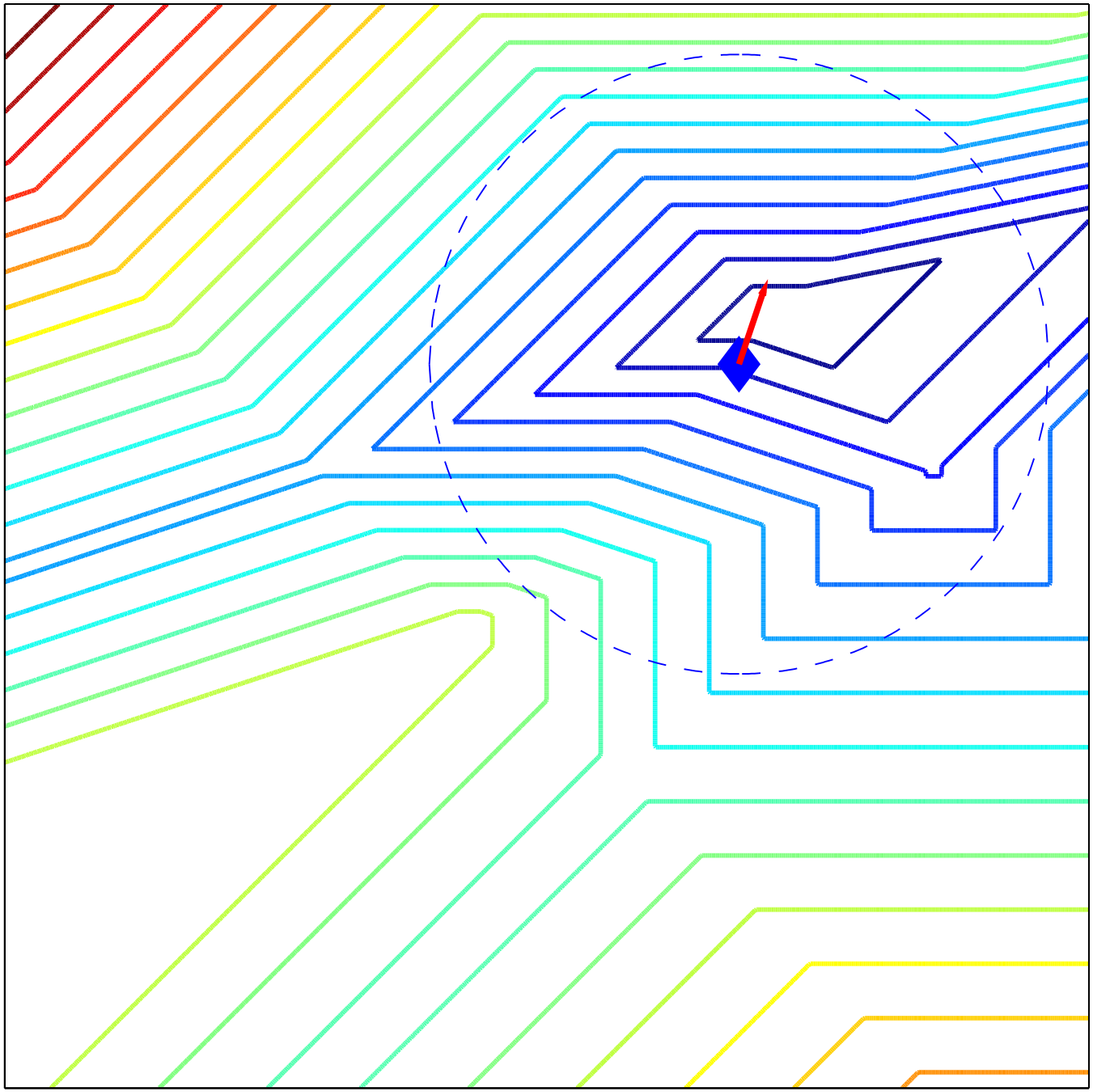


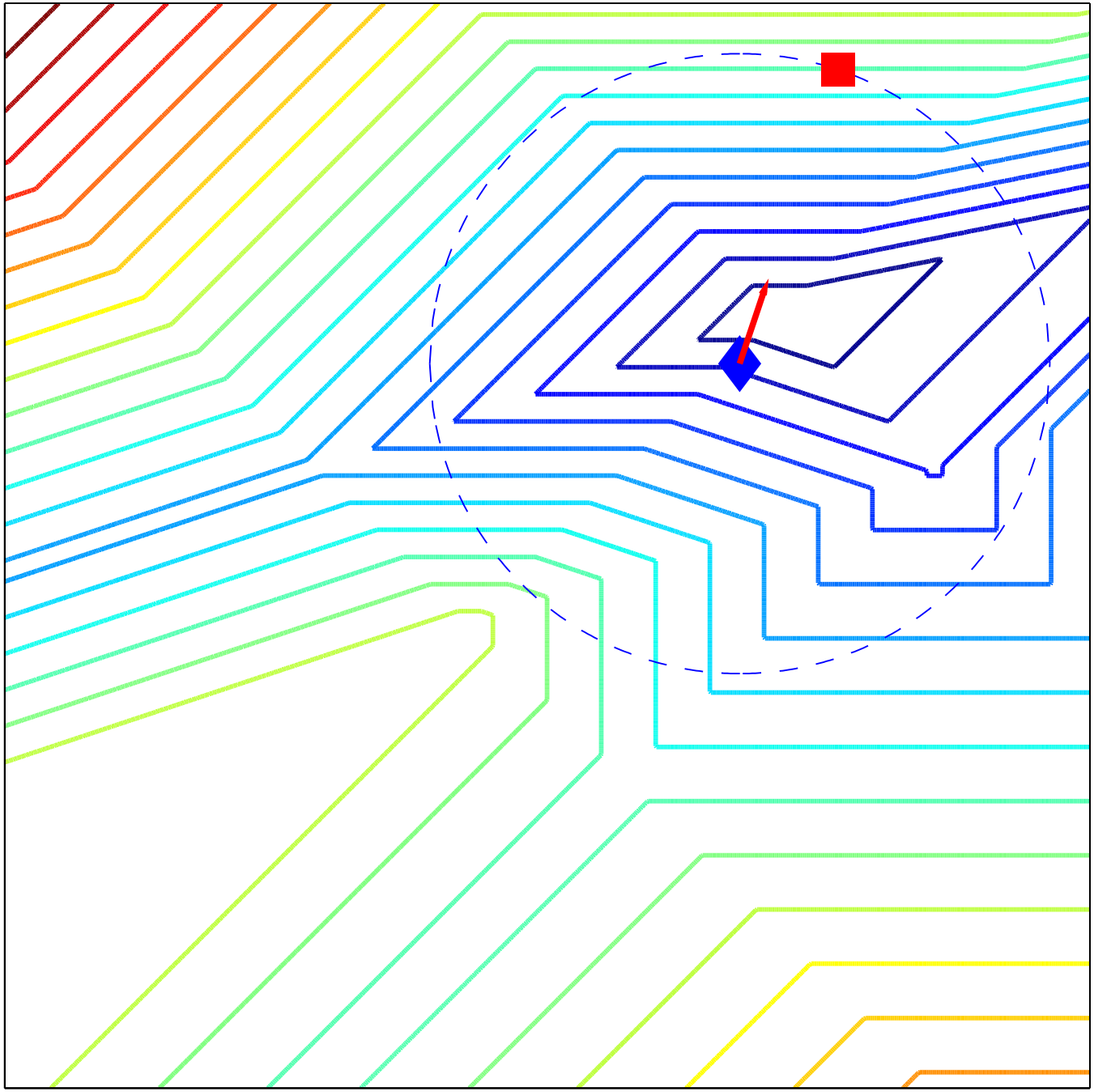


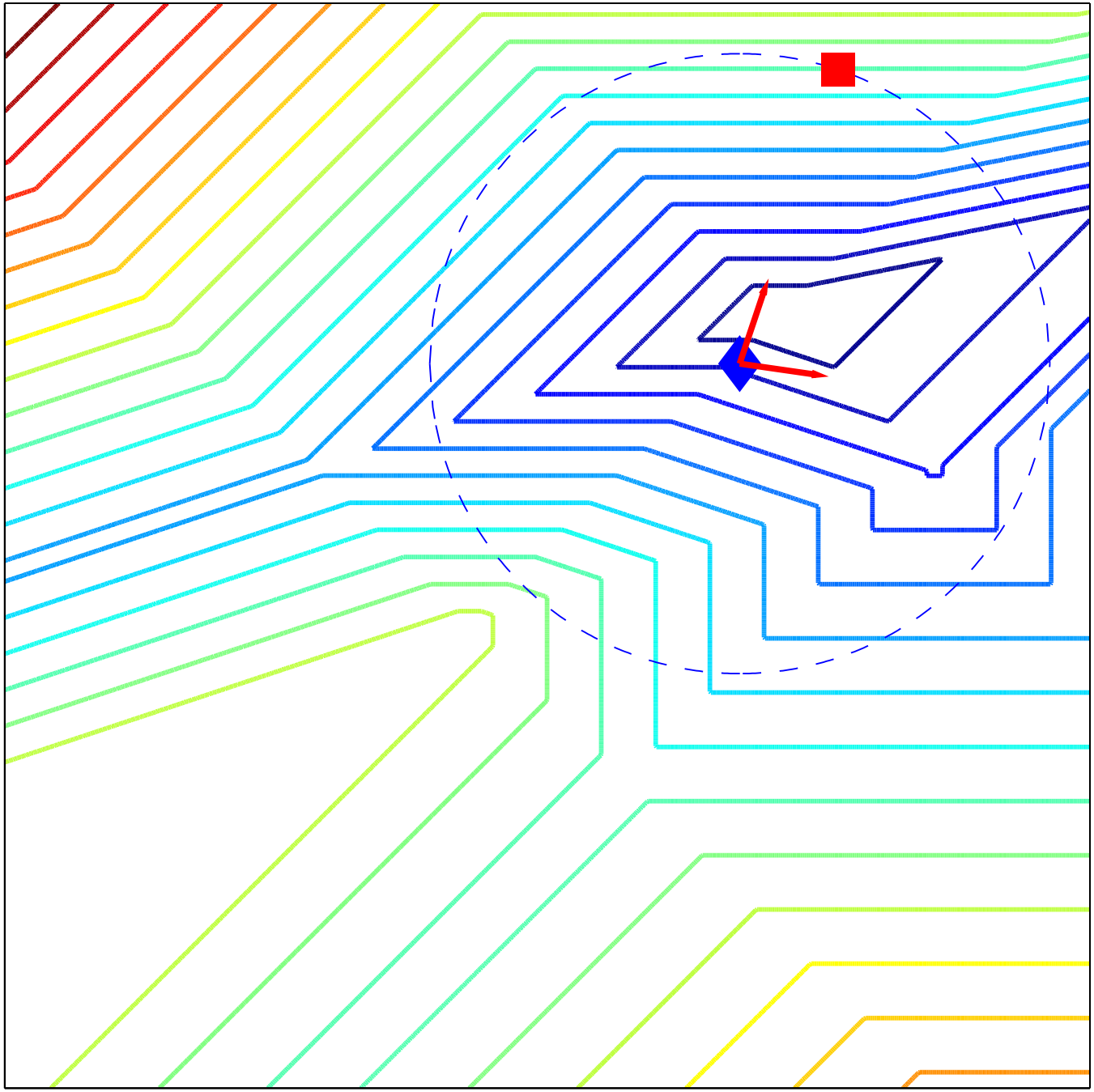


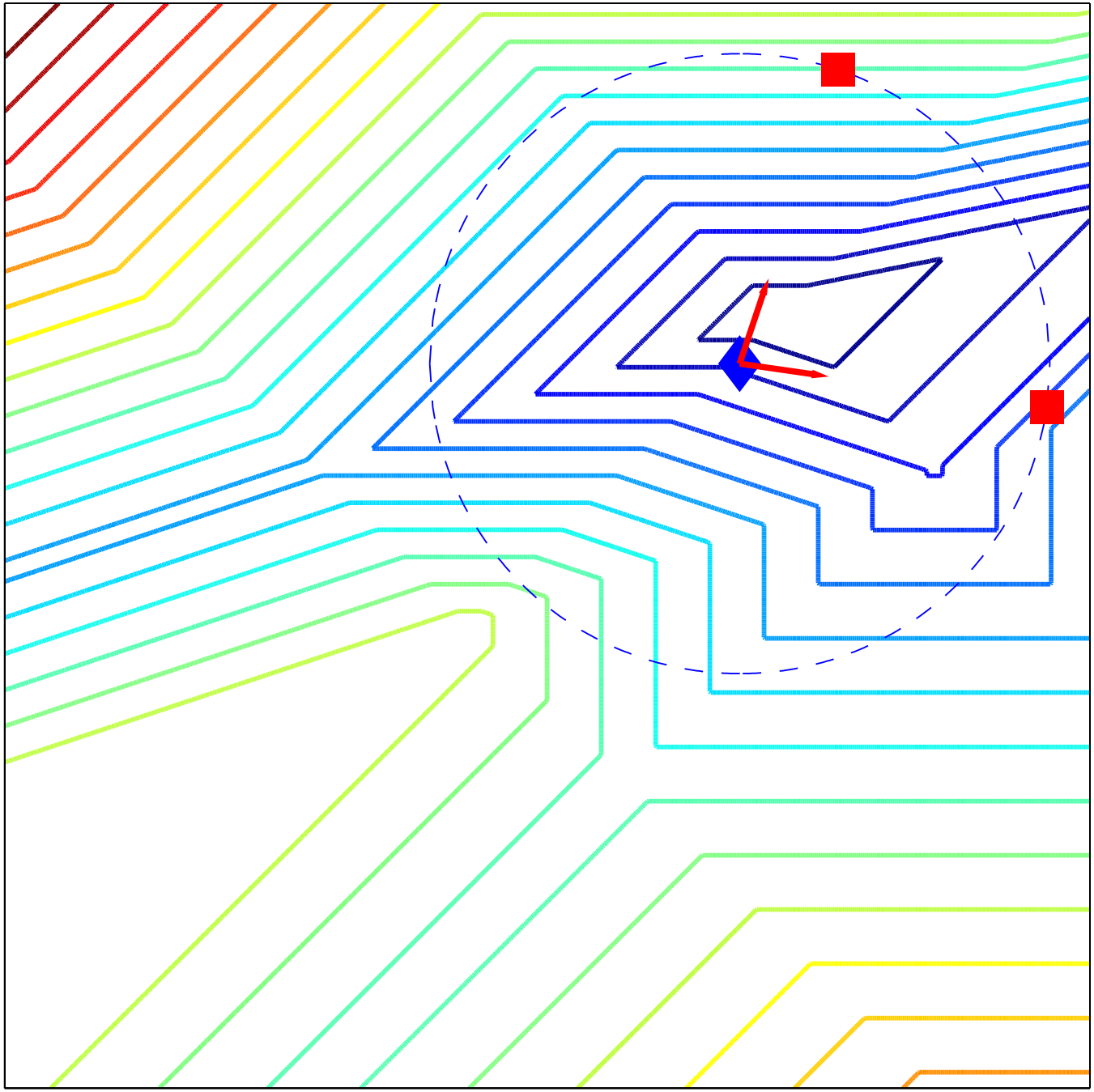


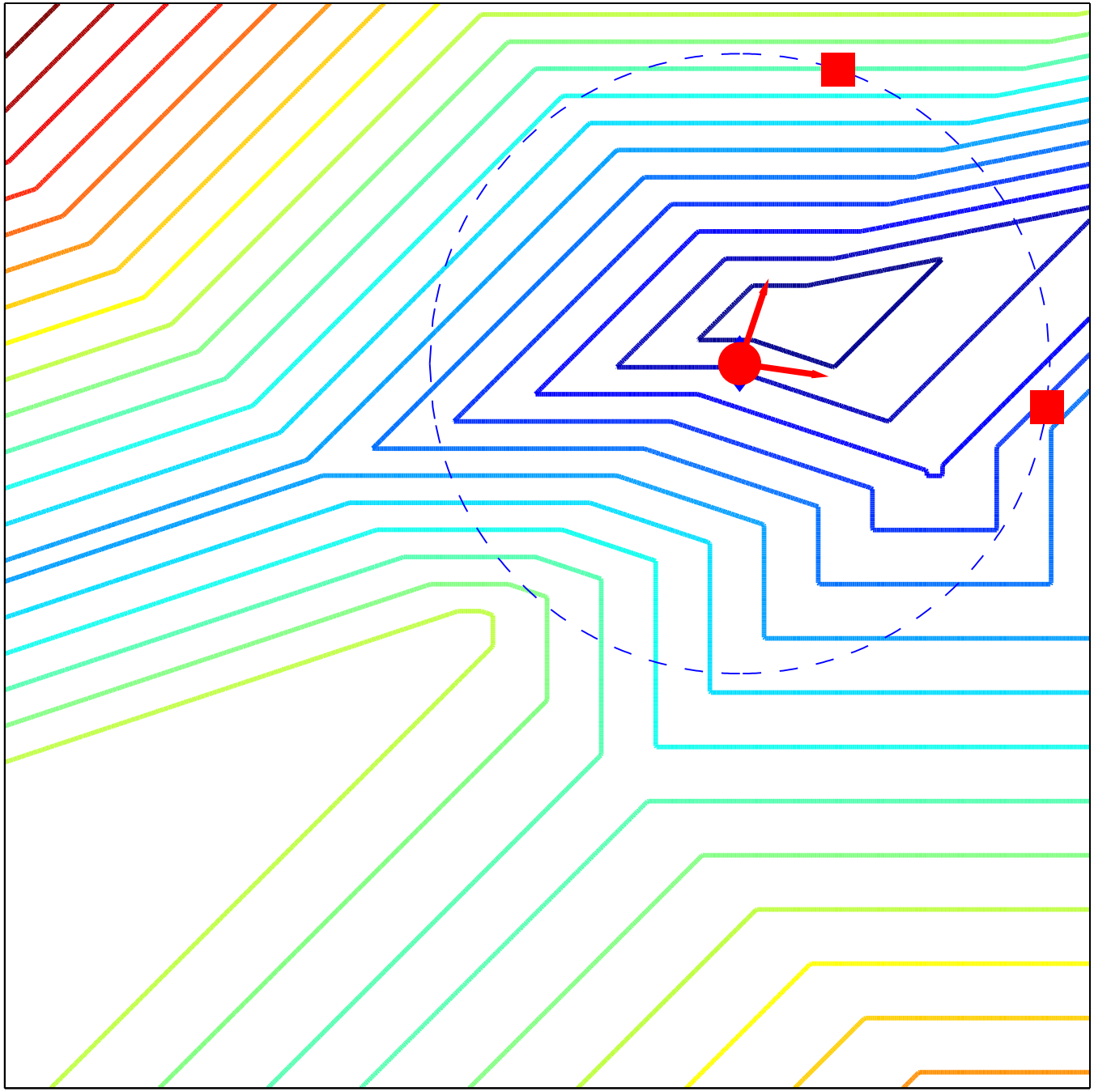


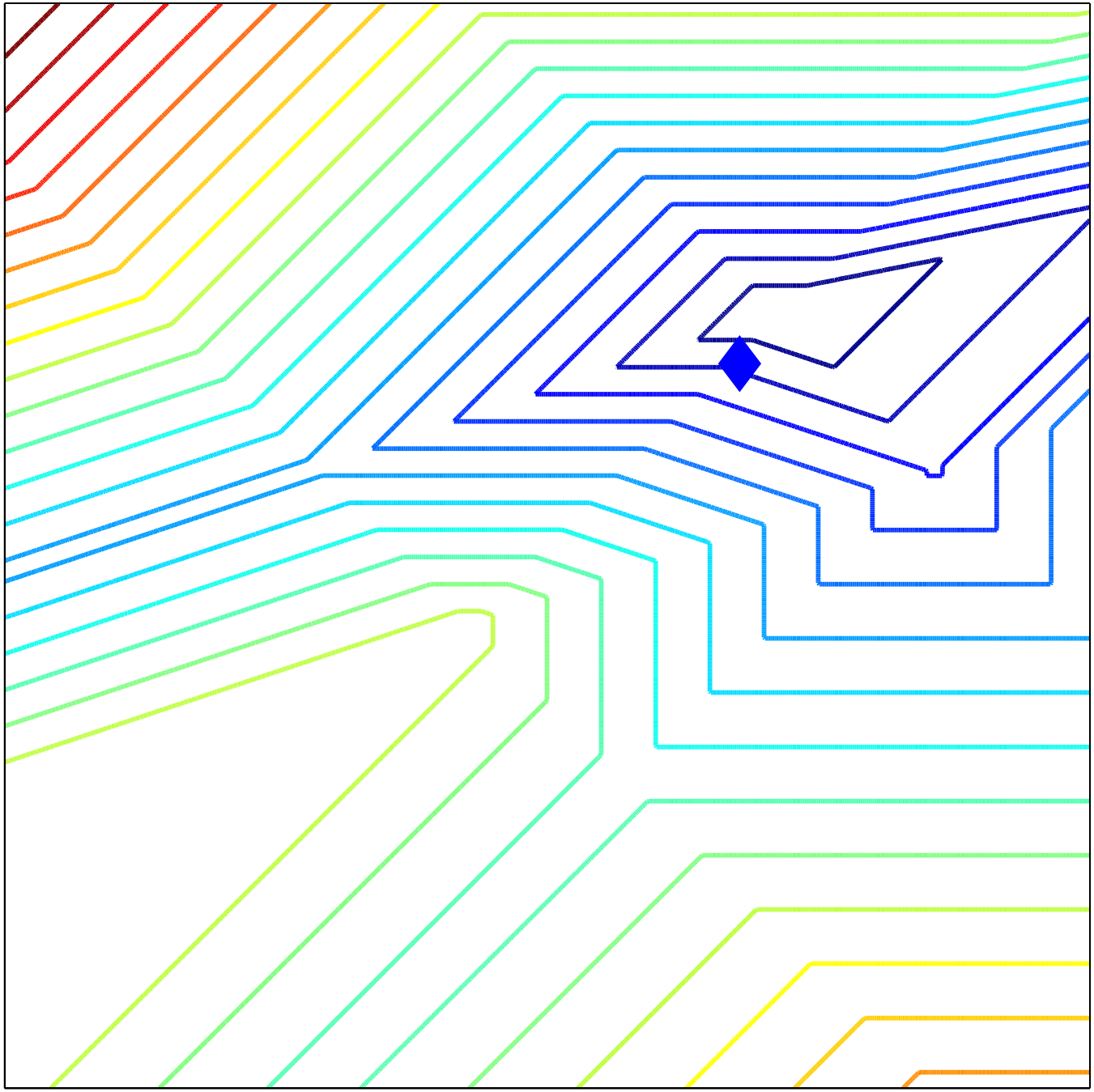


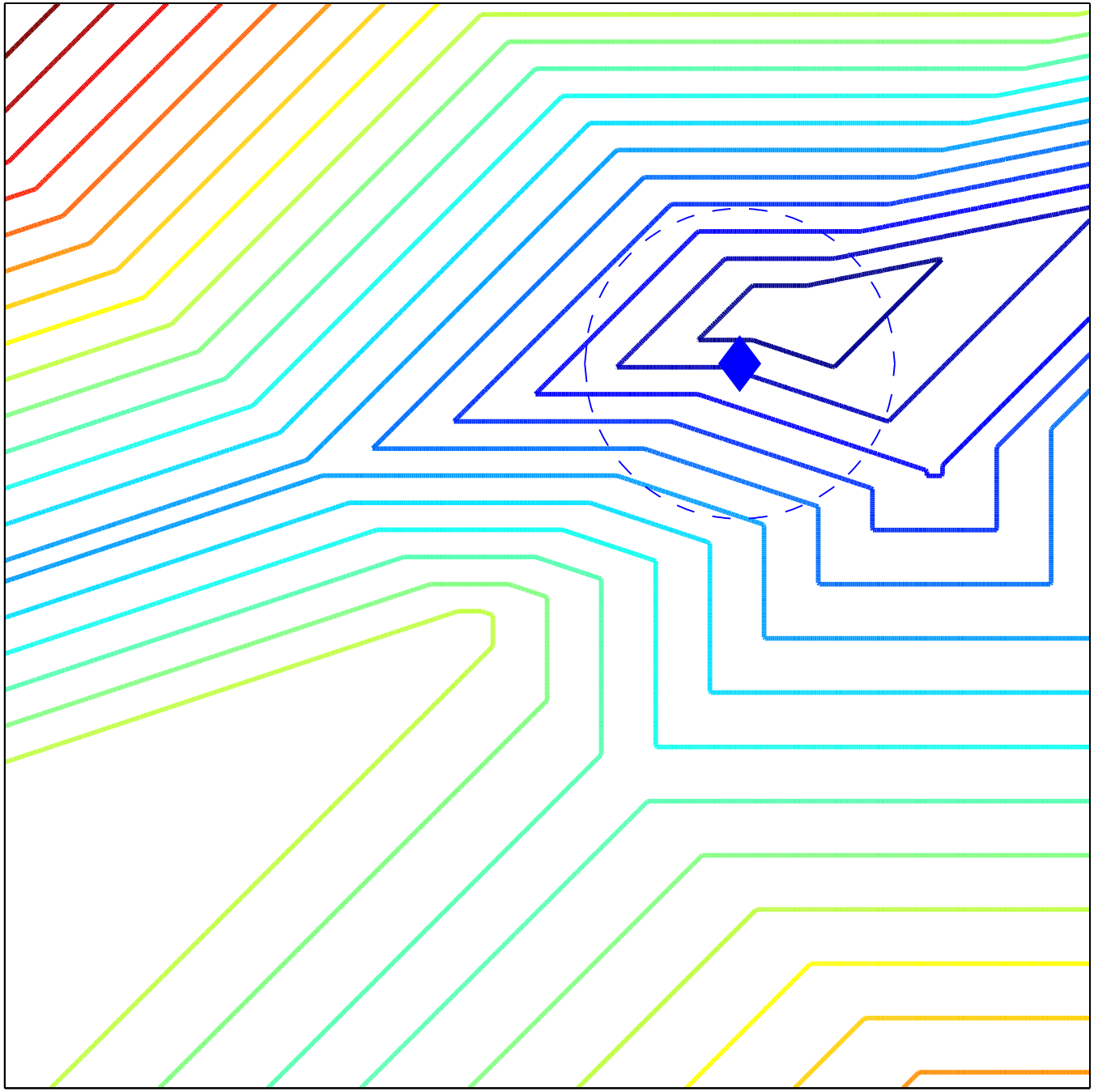


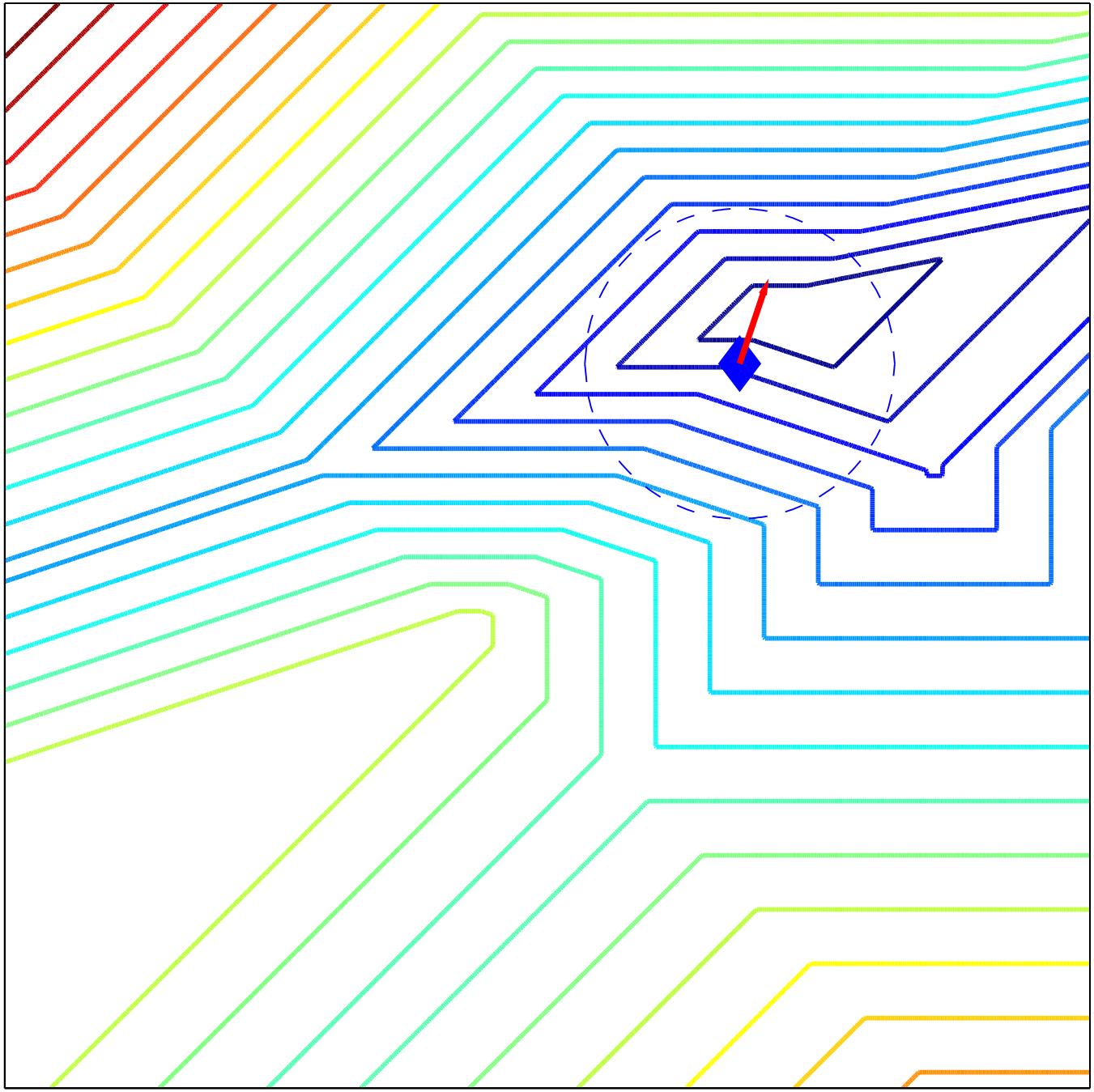


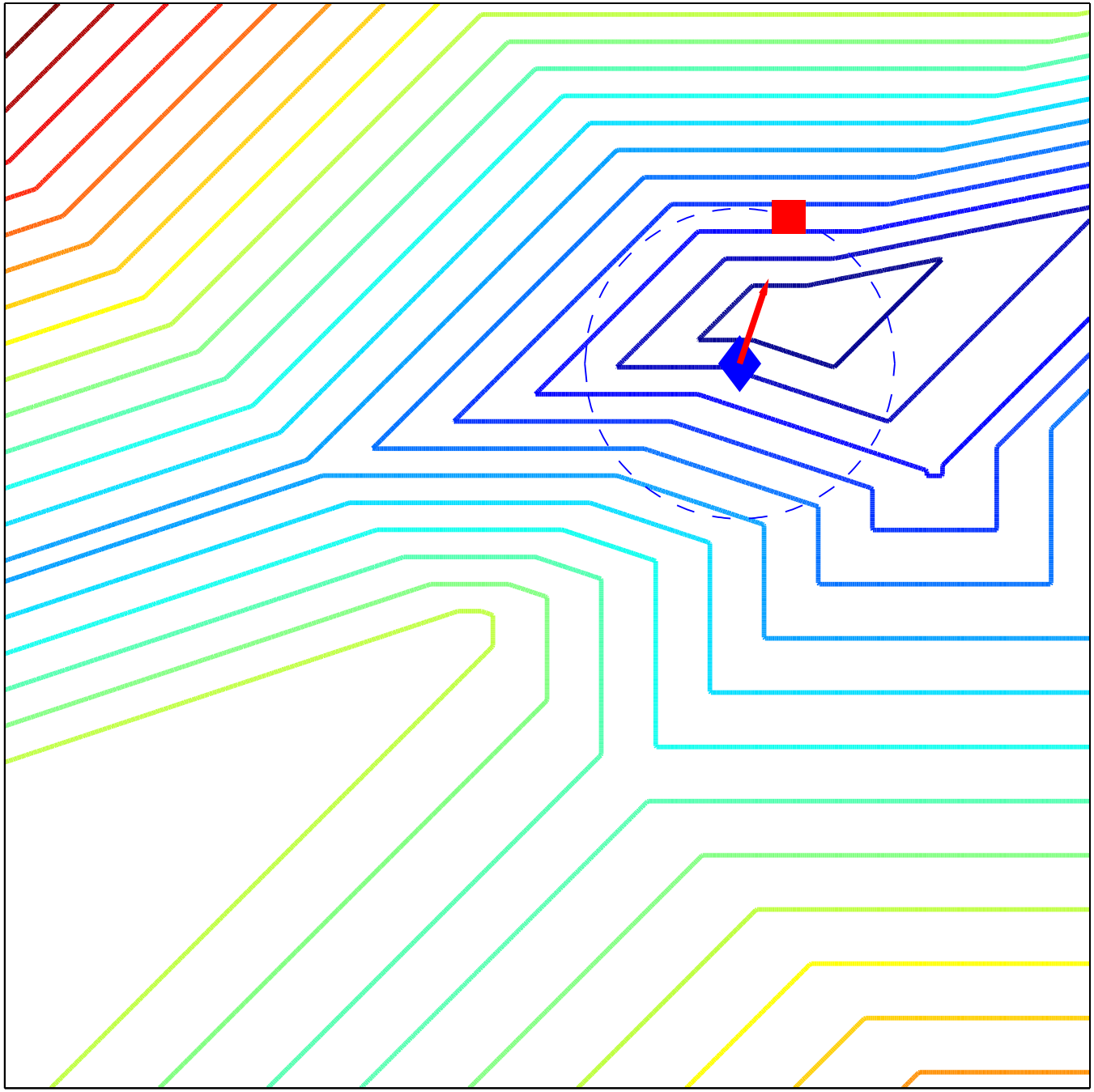


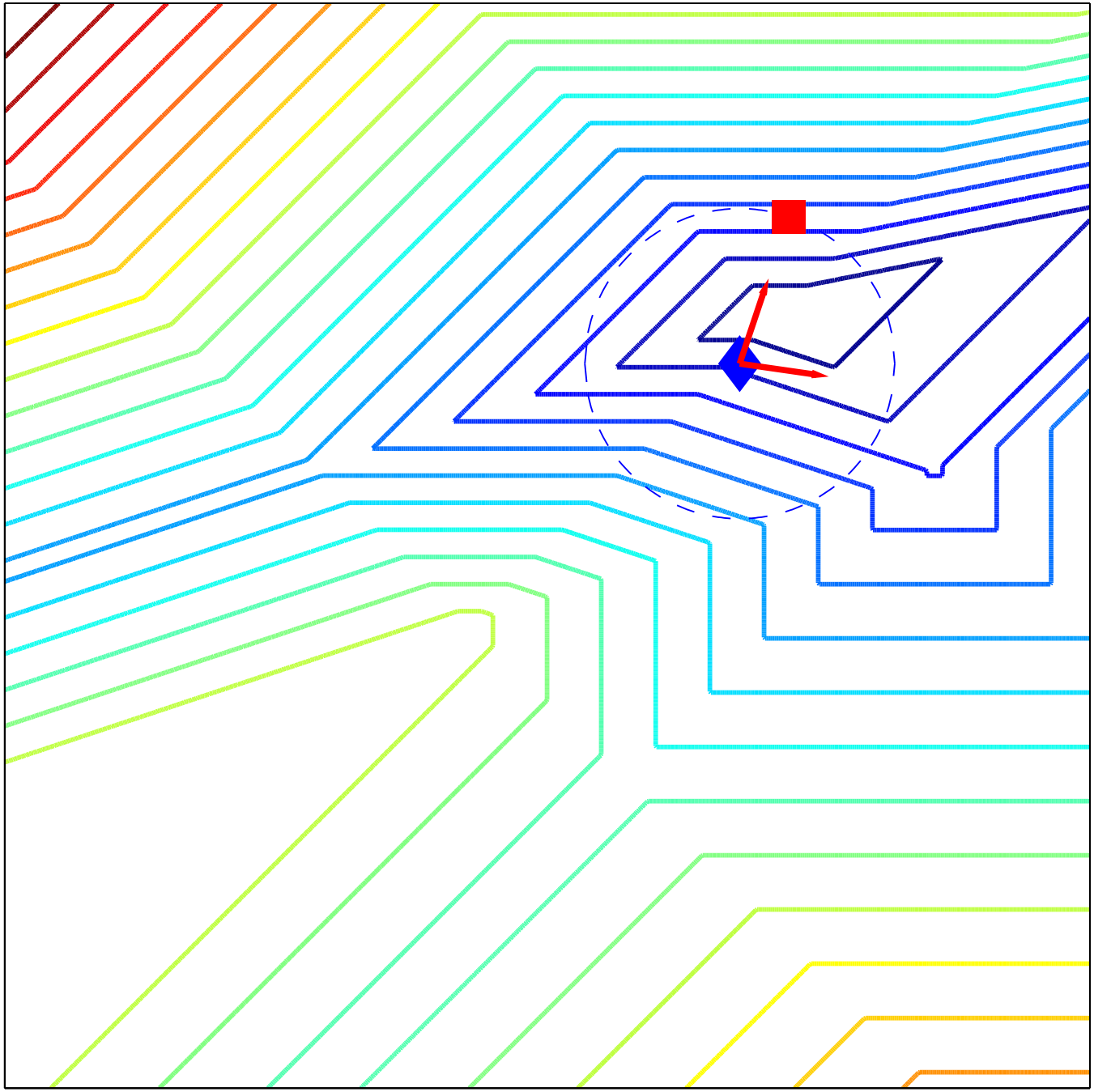


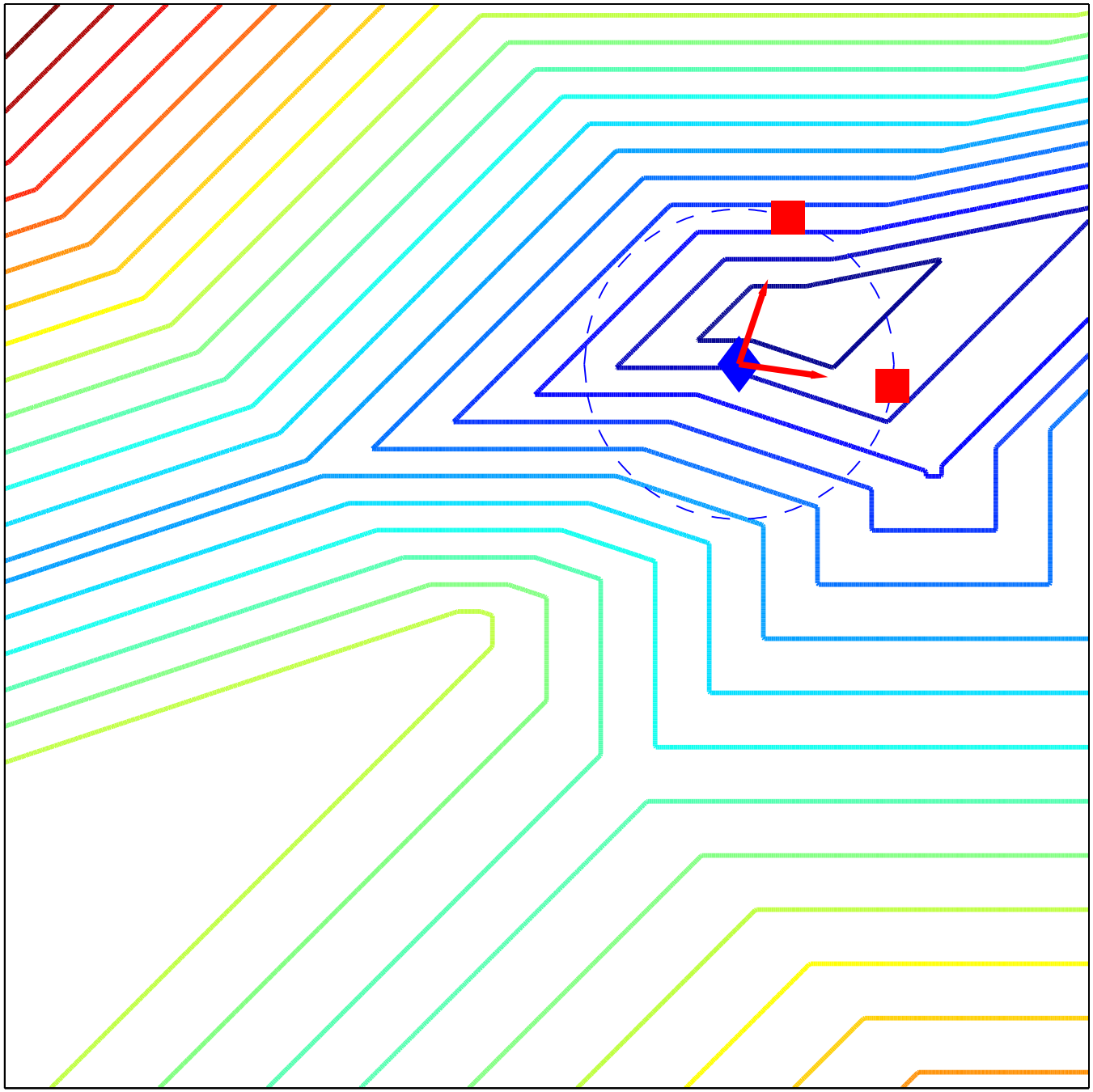


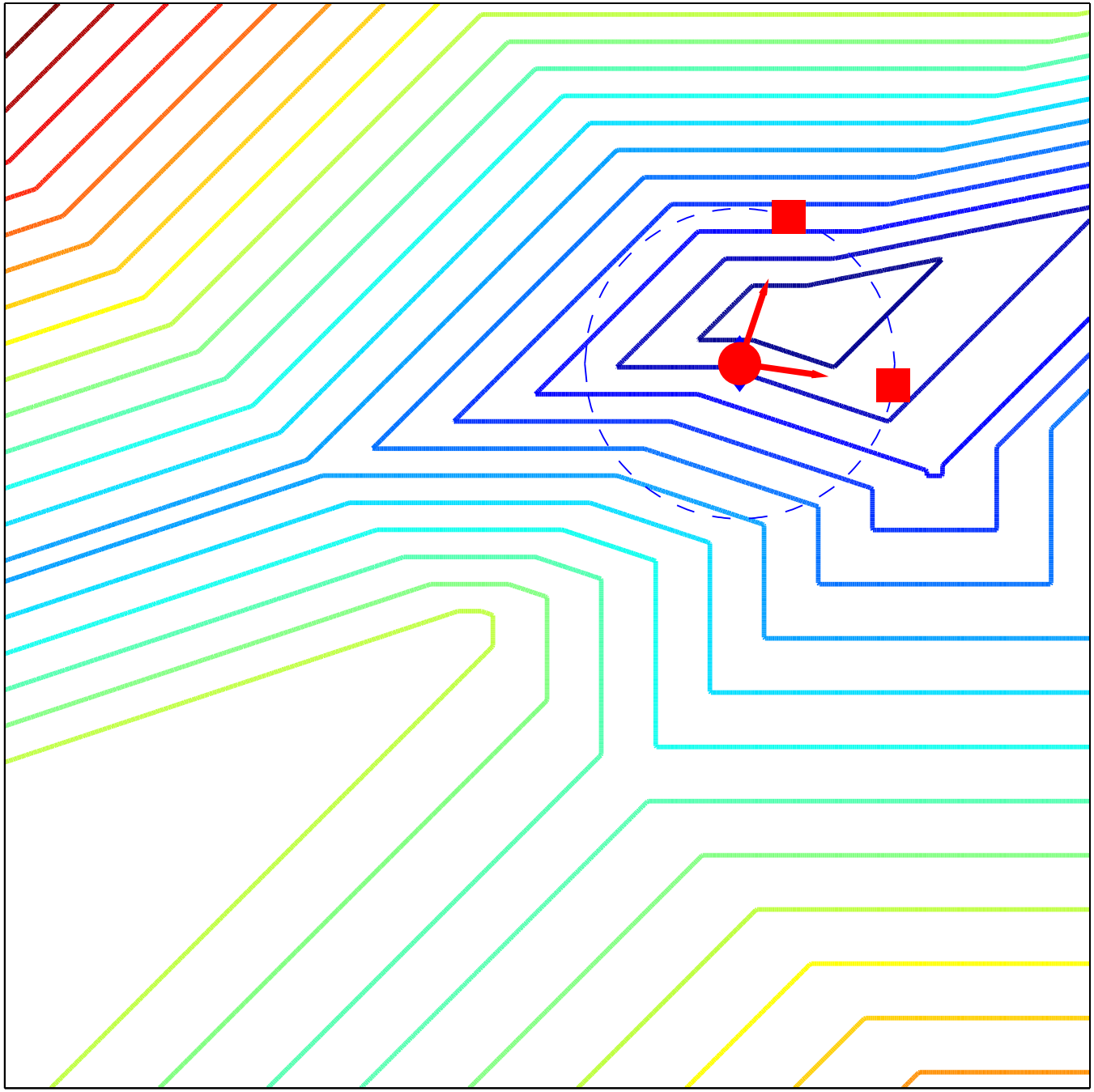


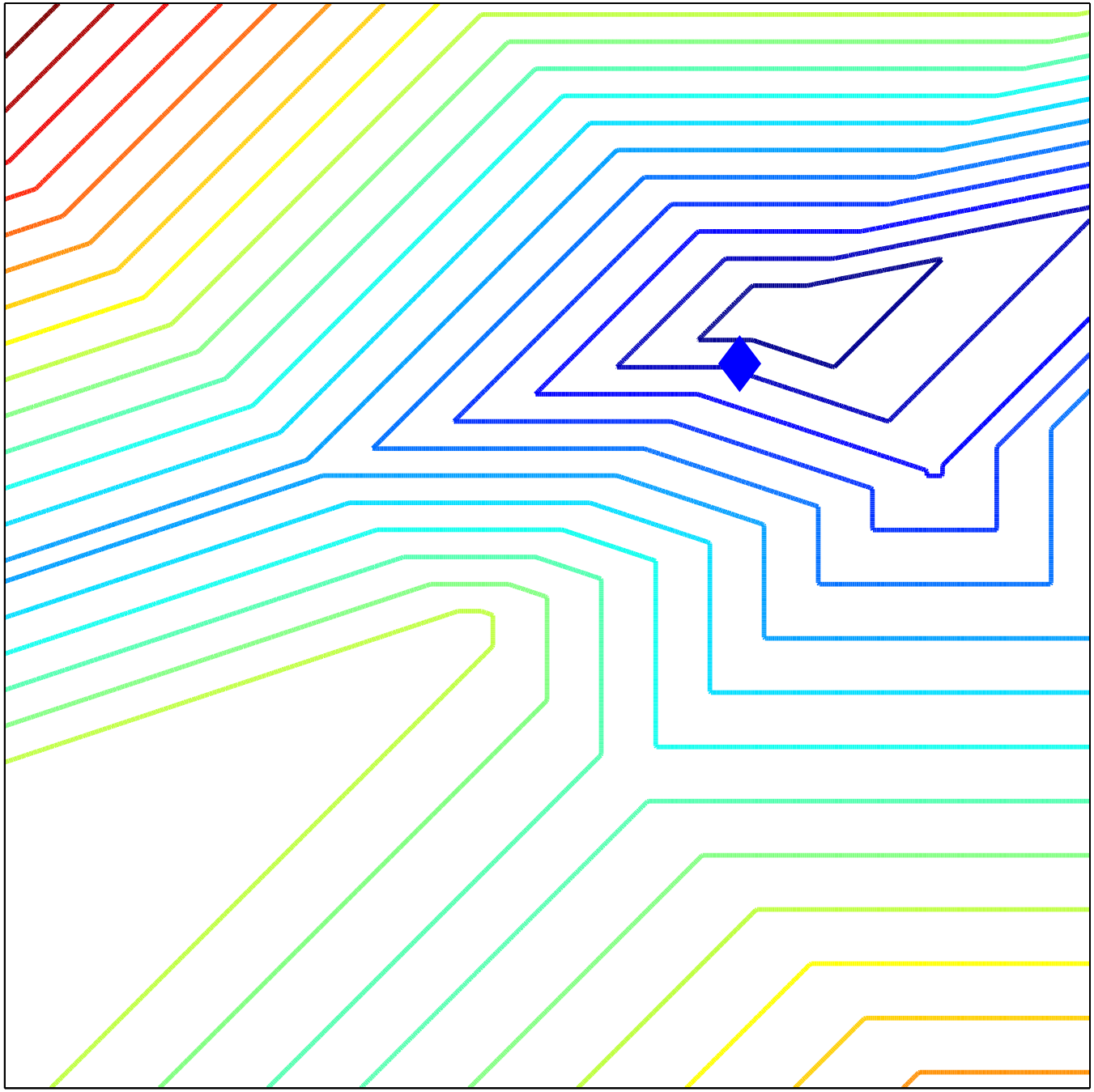


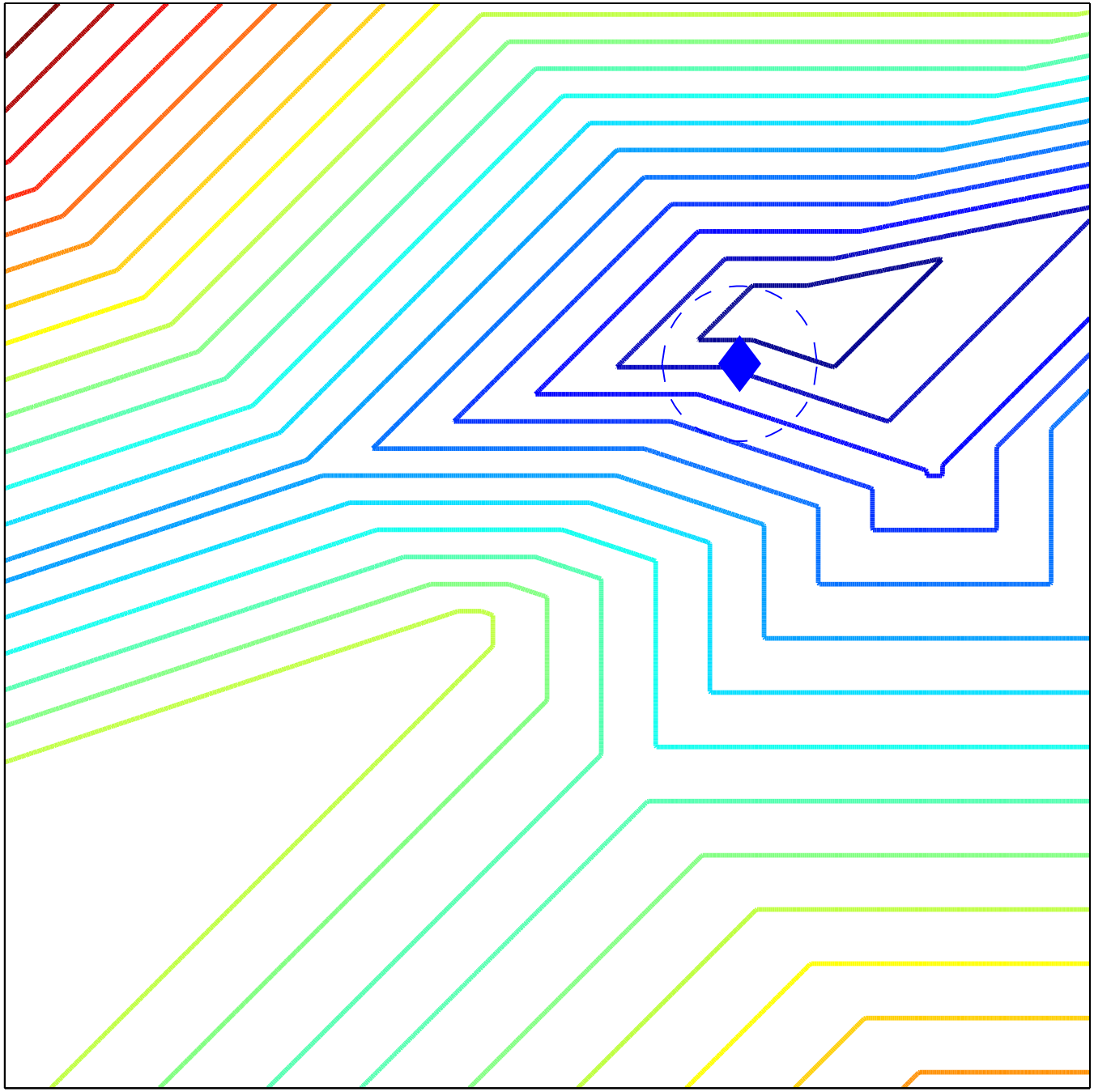


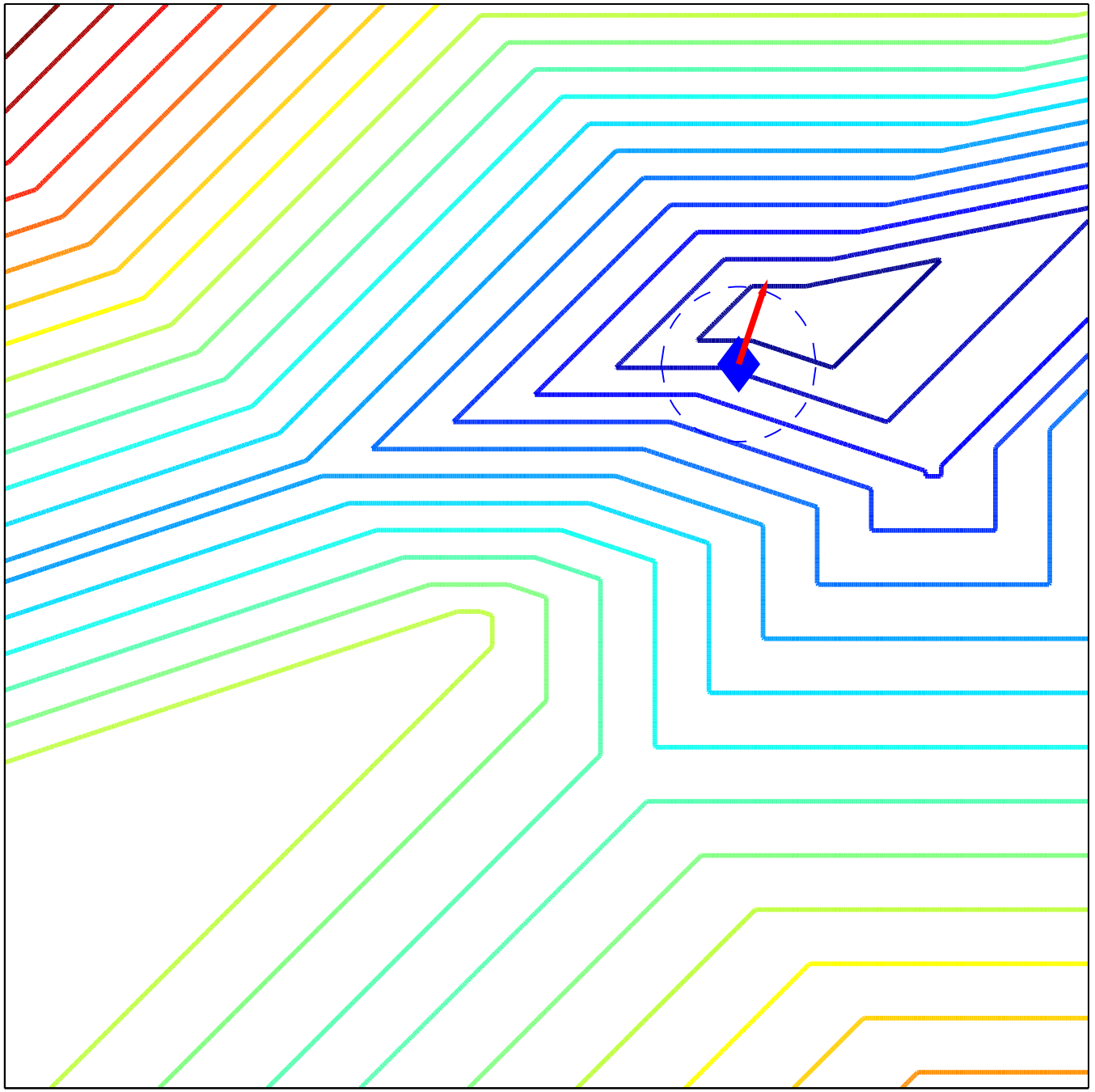


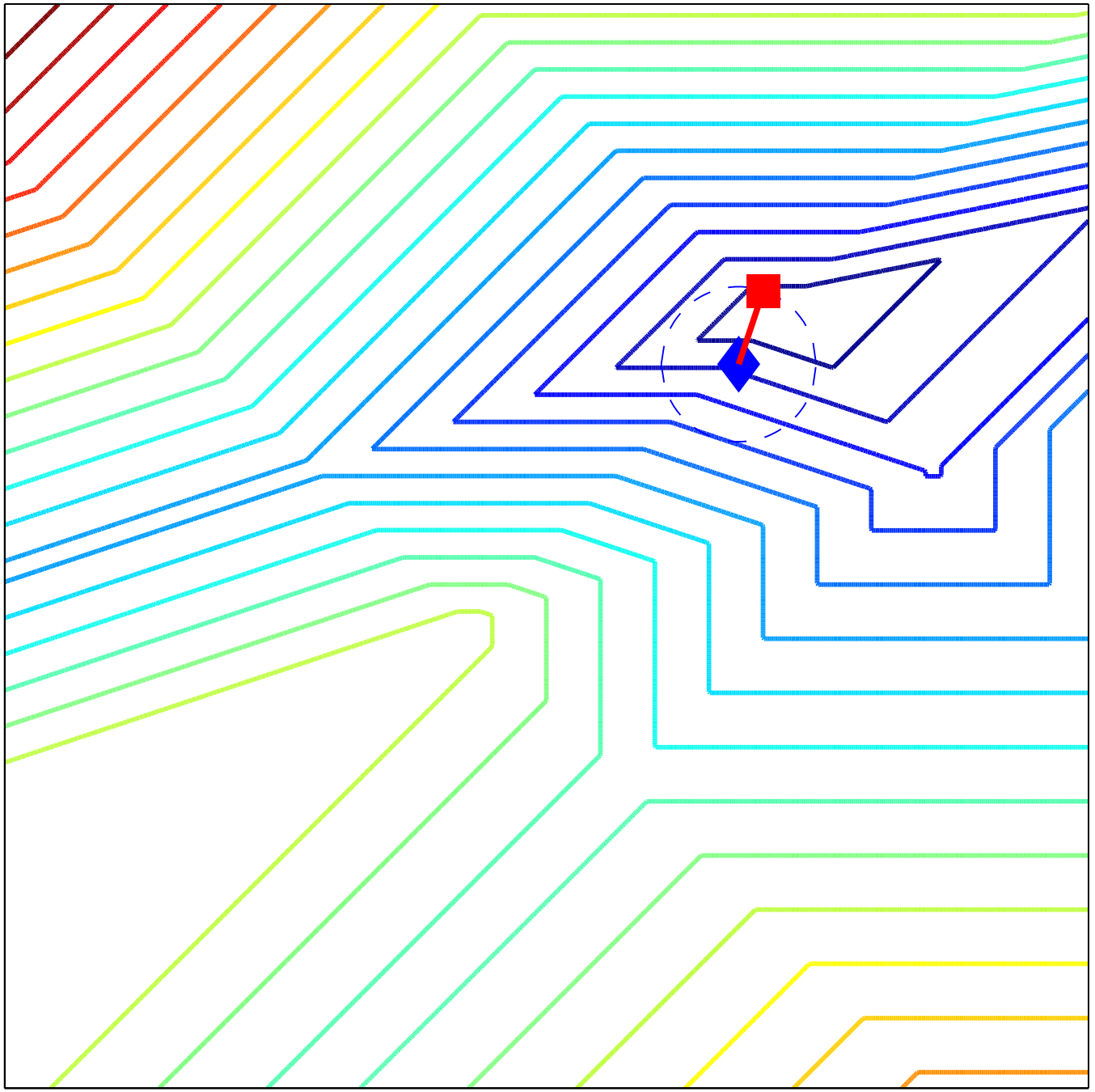


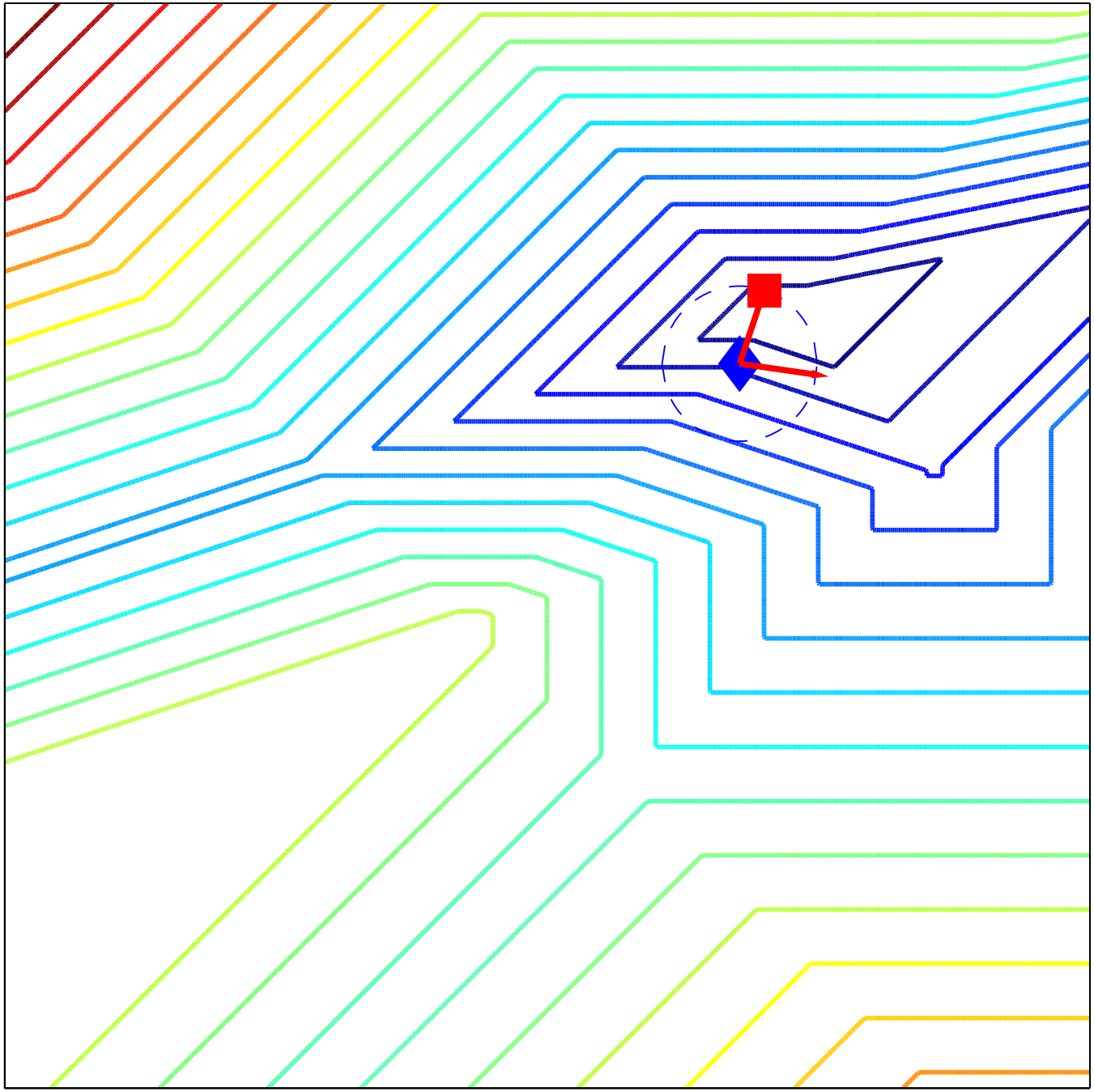


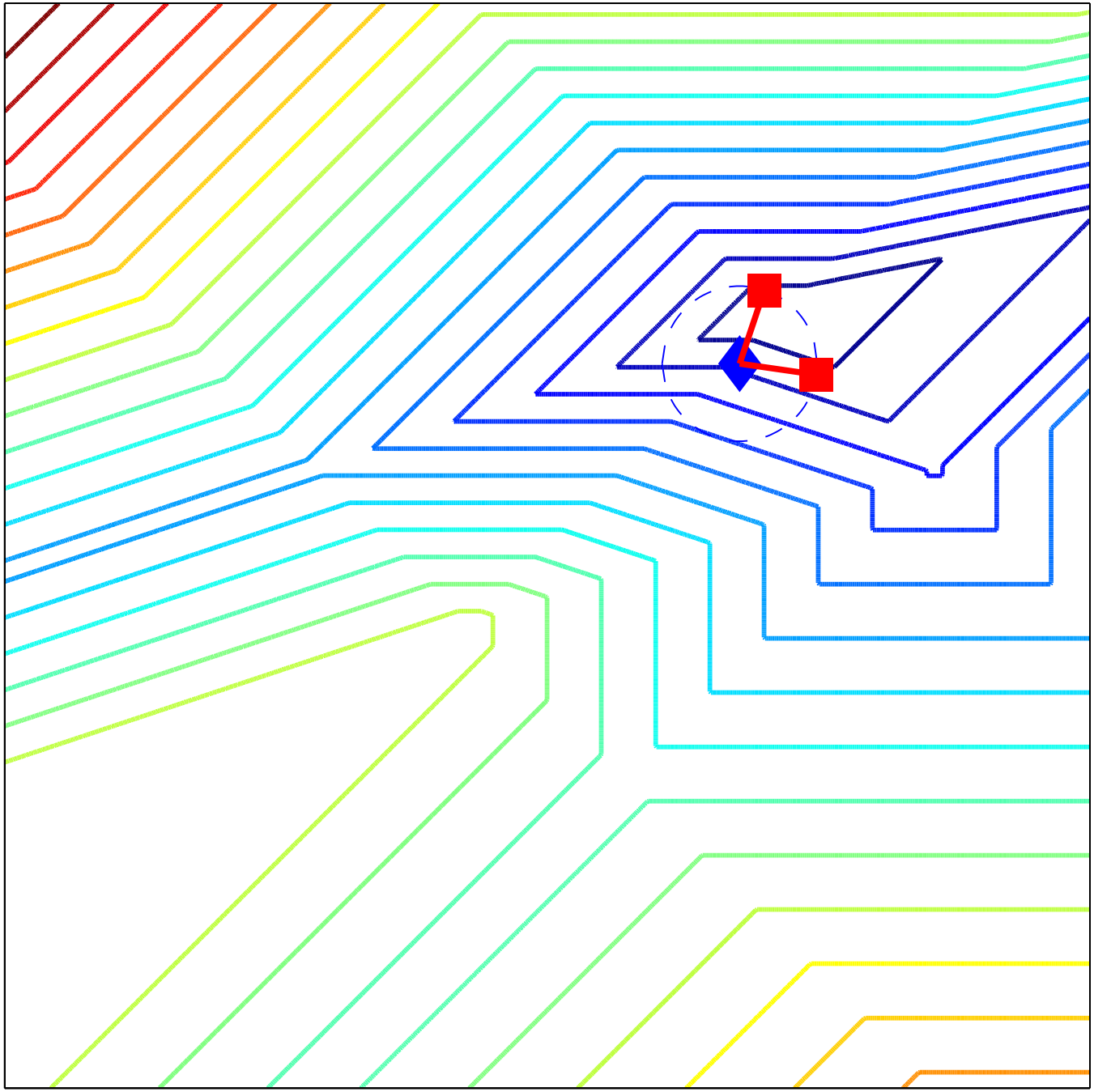


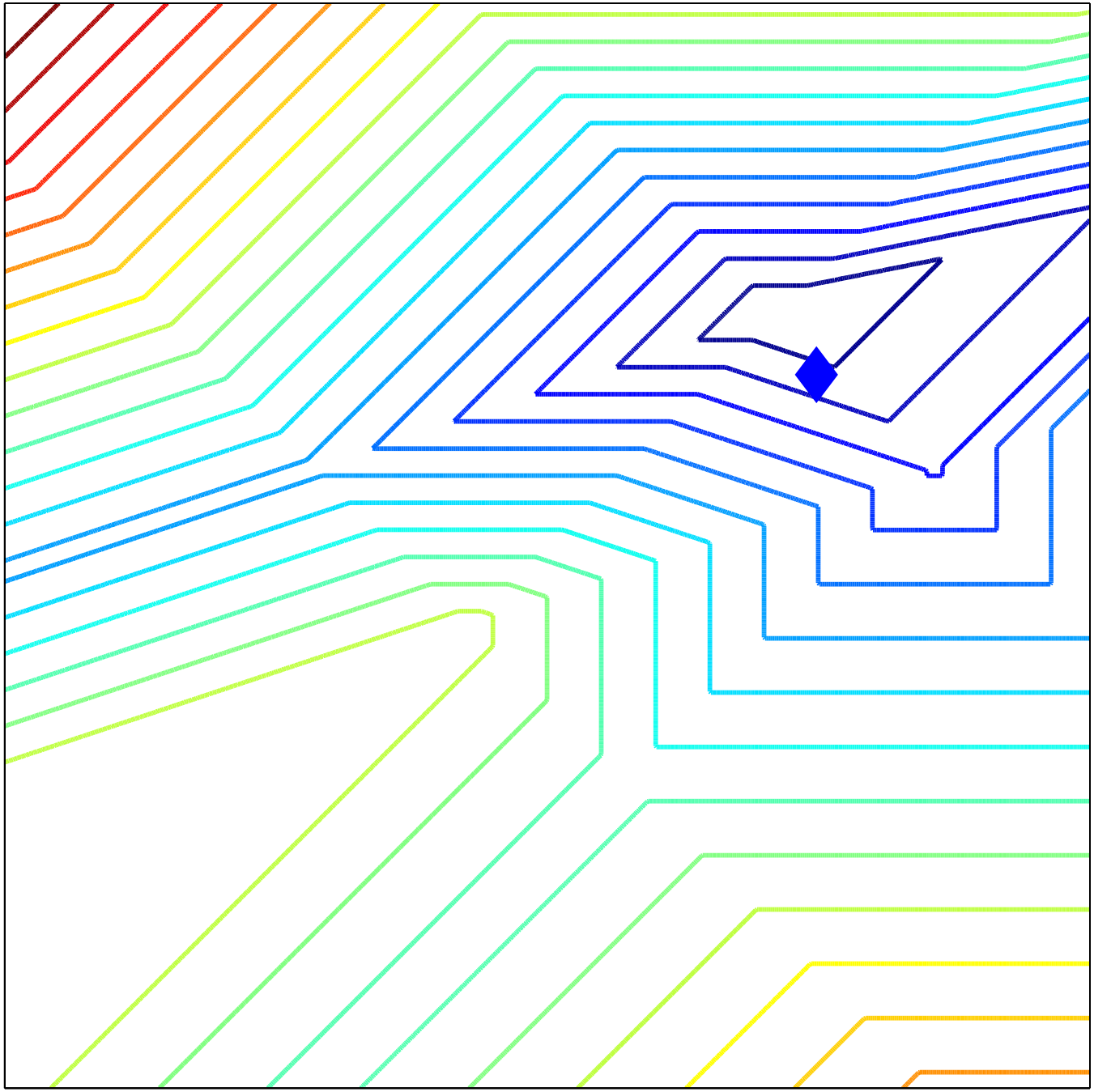












Generator set

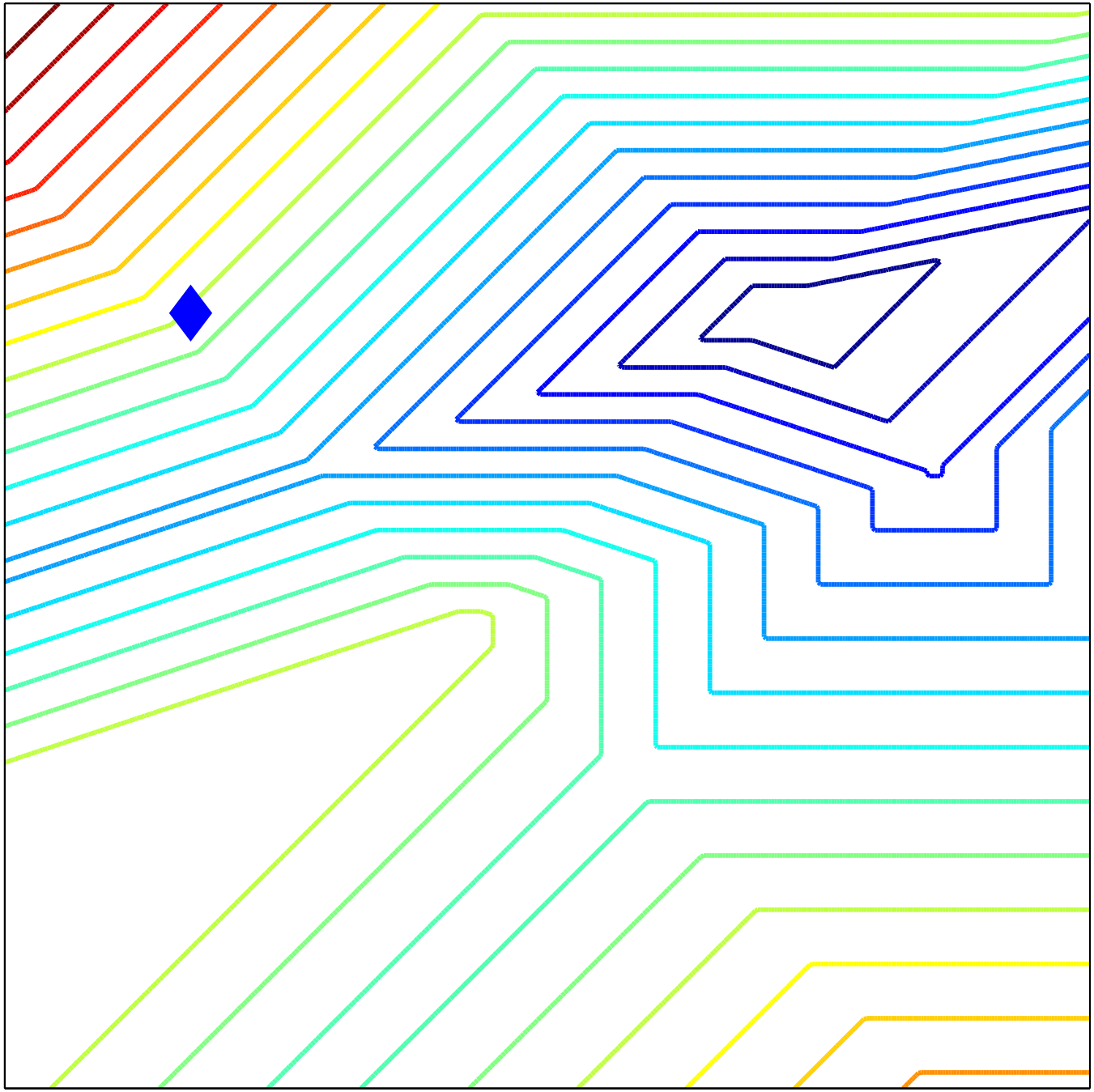
At some iterate x^k,

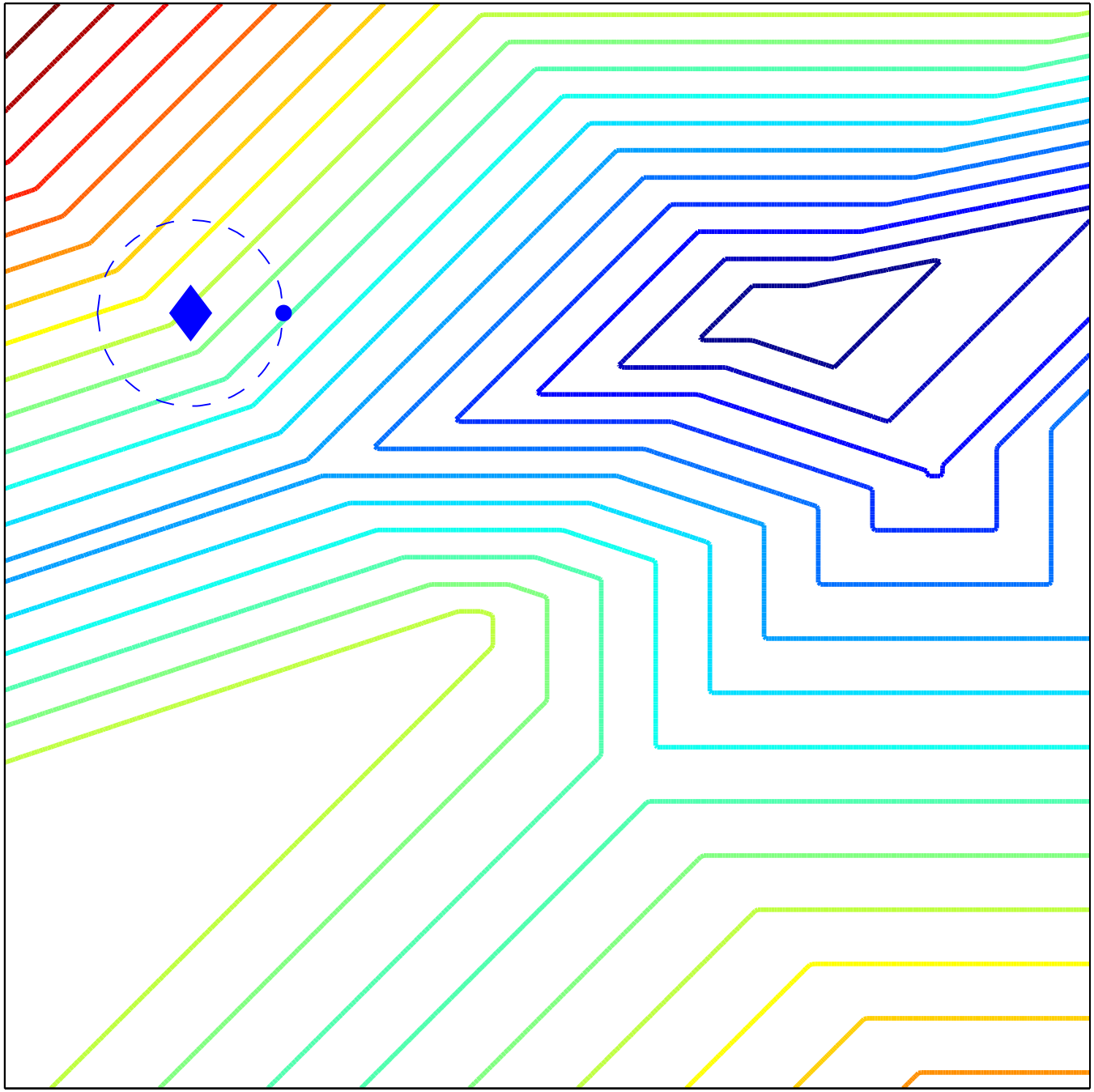
$$\mathfrak{G}^k \triangleq \bigcup_{i \in I_h(F(x^k))} \left\{ \nabla \psi(x^k) + \nabla M(x^k) a_i \right\} \rightarrow \text{MS4PL-1}$$

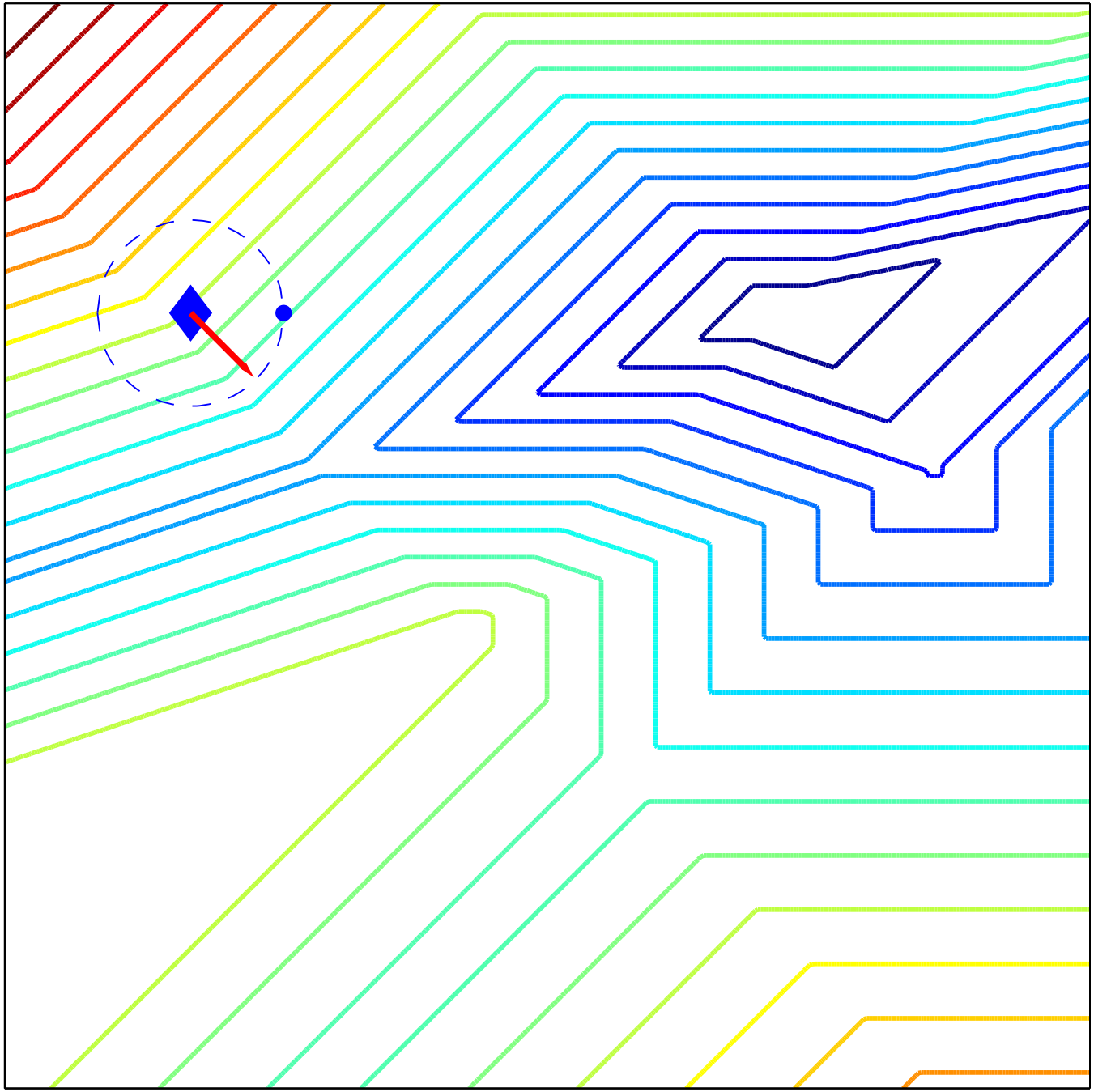
where $I_h(F(x^k))$ is the set of essentially active indices of h at $F(x^k)$.

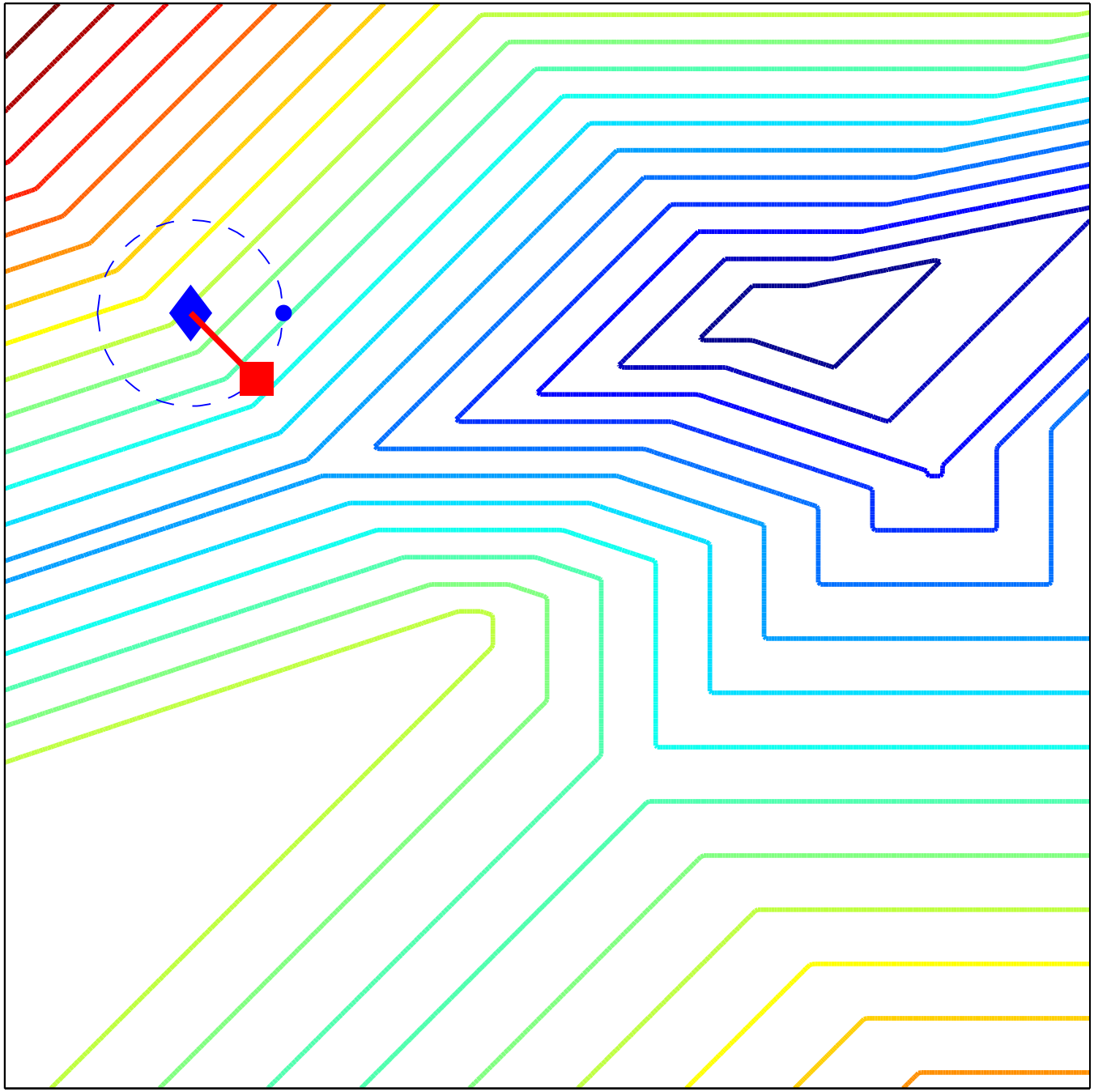
Or, given a set of points $Y = \left\{ x^k, y^2, \ldots, y^p \right\} \subset \mathcal{B}(x^k, \Delta_k)$,

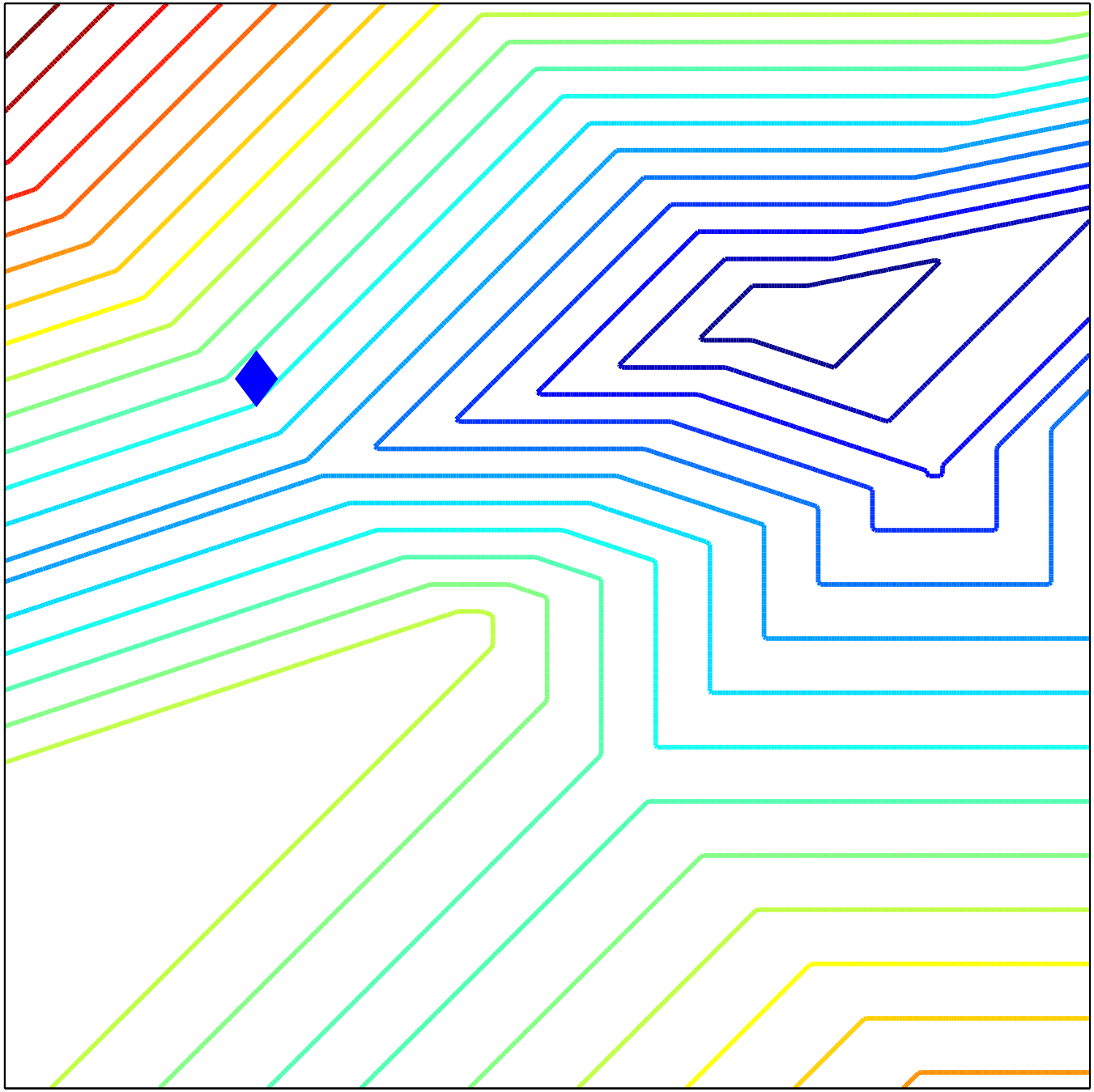
$$\mathfrak{G}^k \triangleq \bigcup_{y \in Y} \bigcup_{i \in I_h(F(y))} \left\{ \nabla \psi(x^k) + \nabla M(x^k) a_i \right\} \rightarrow \text{MS4PL-2}$$

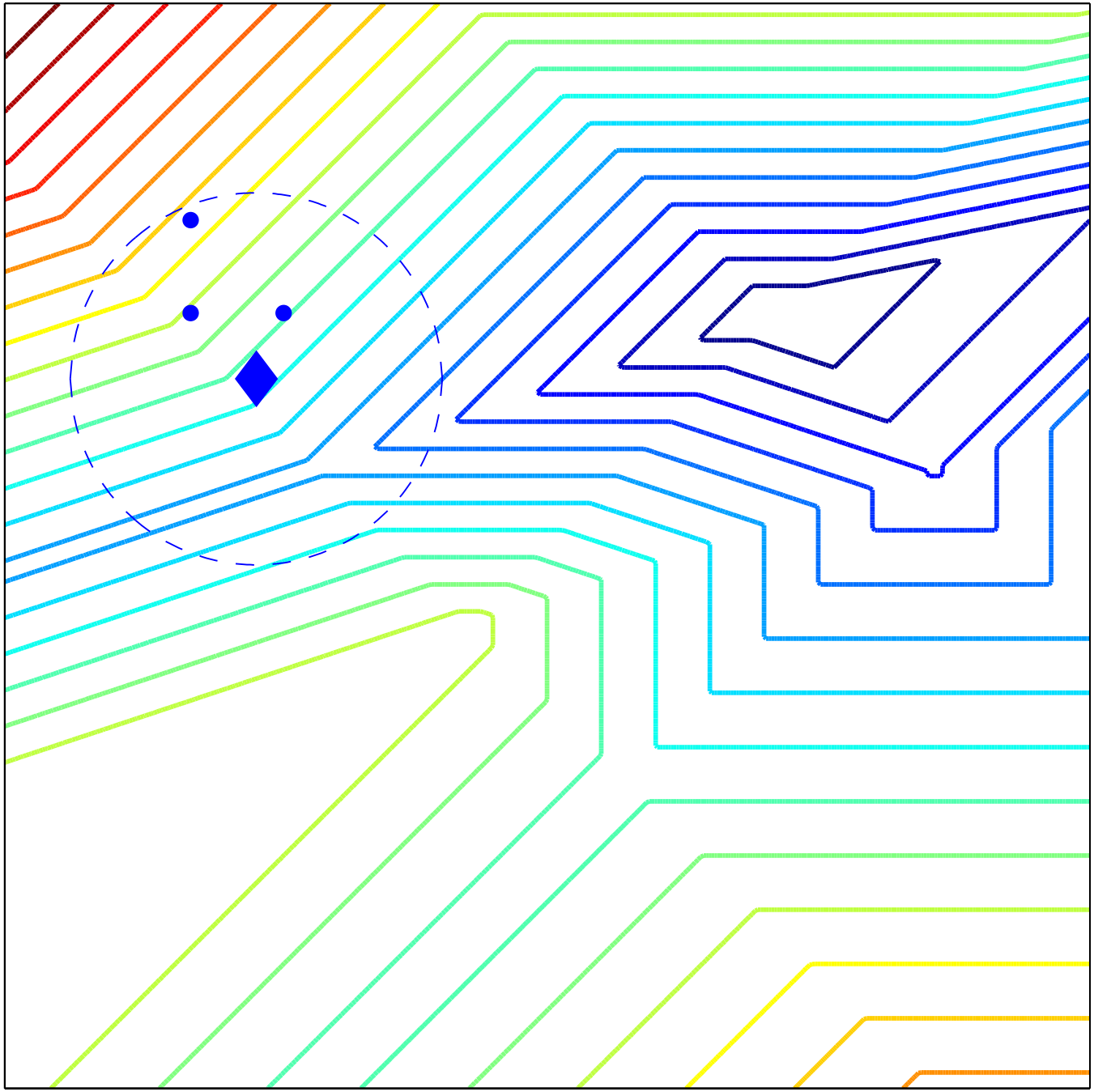


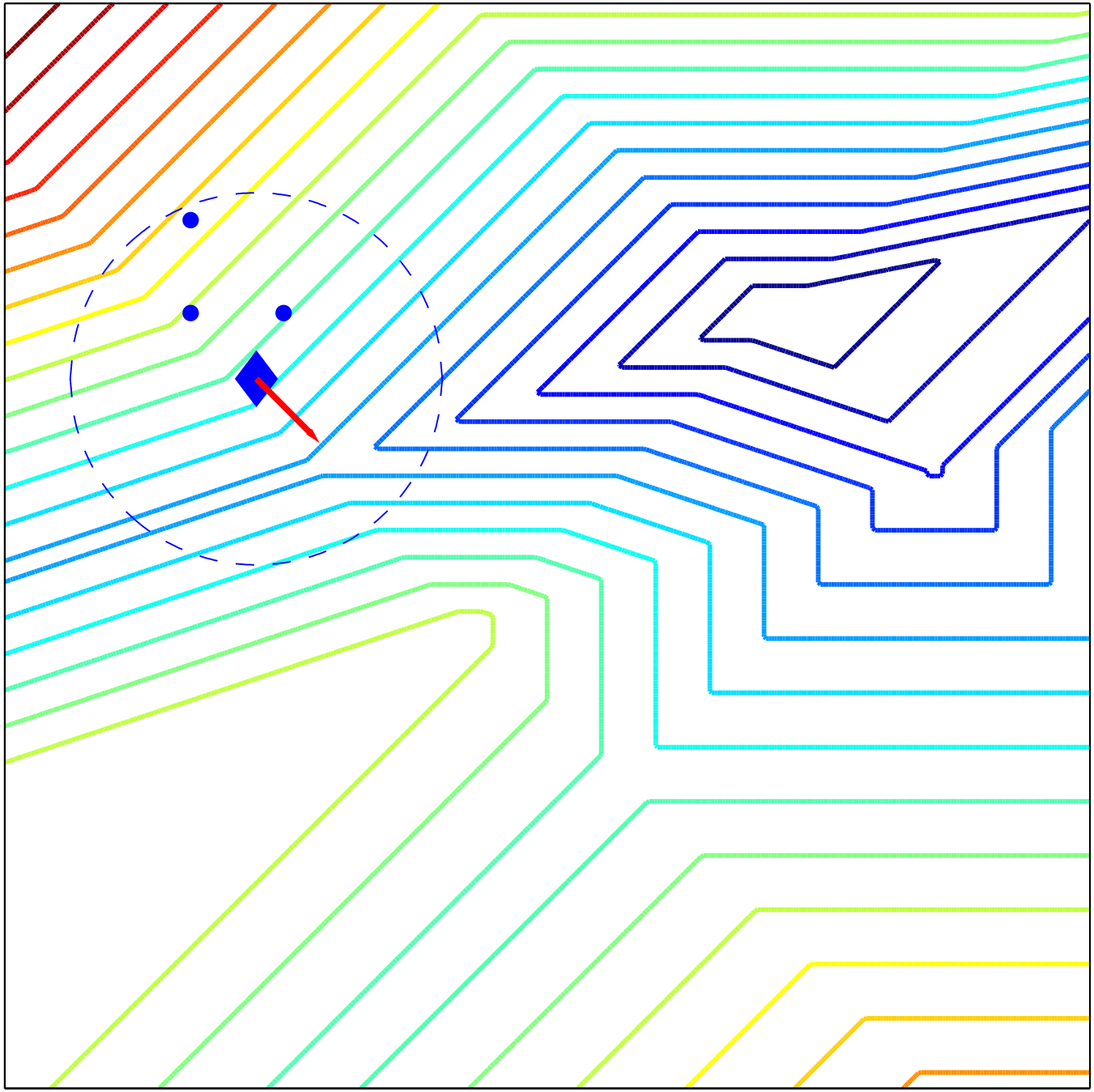


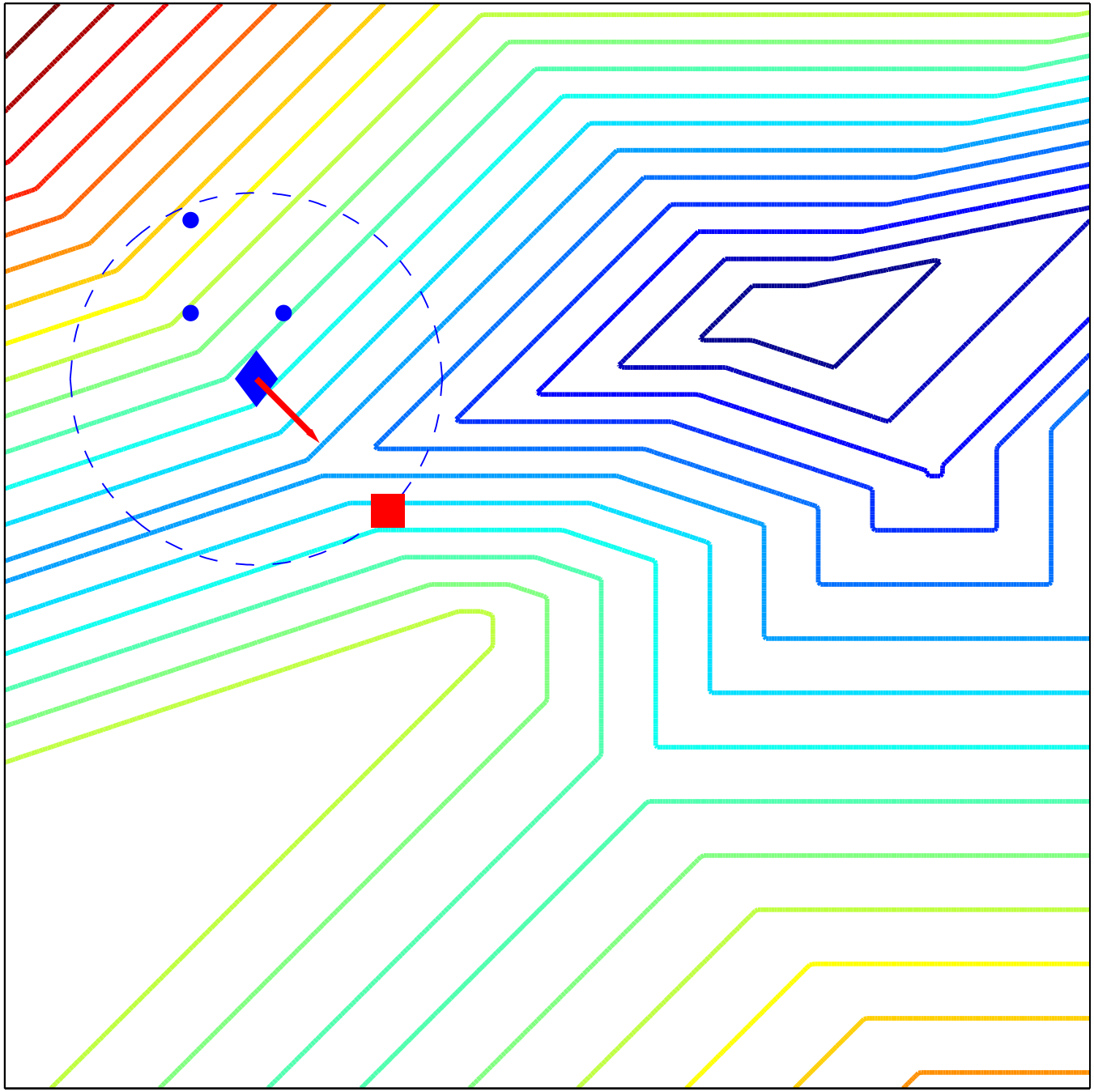


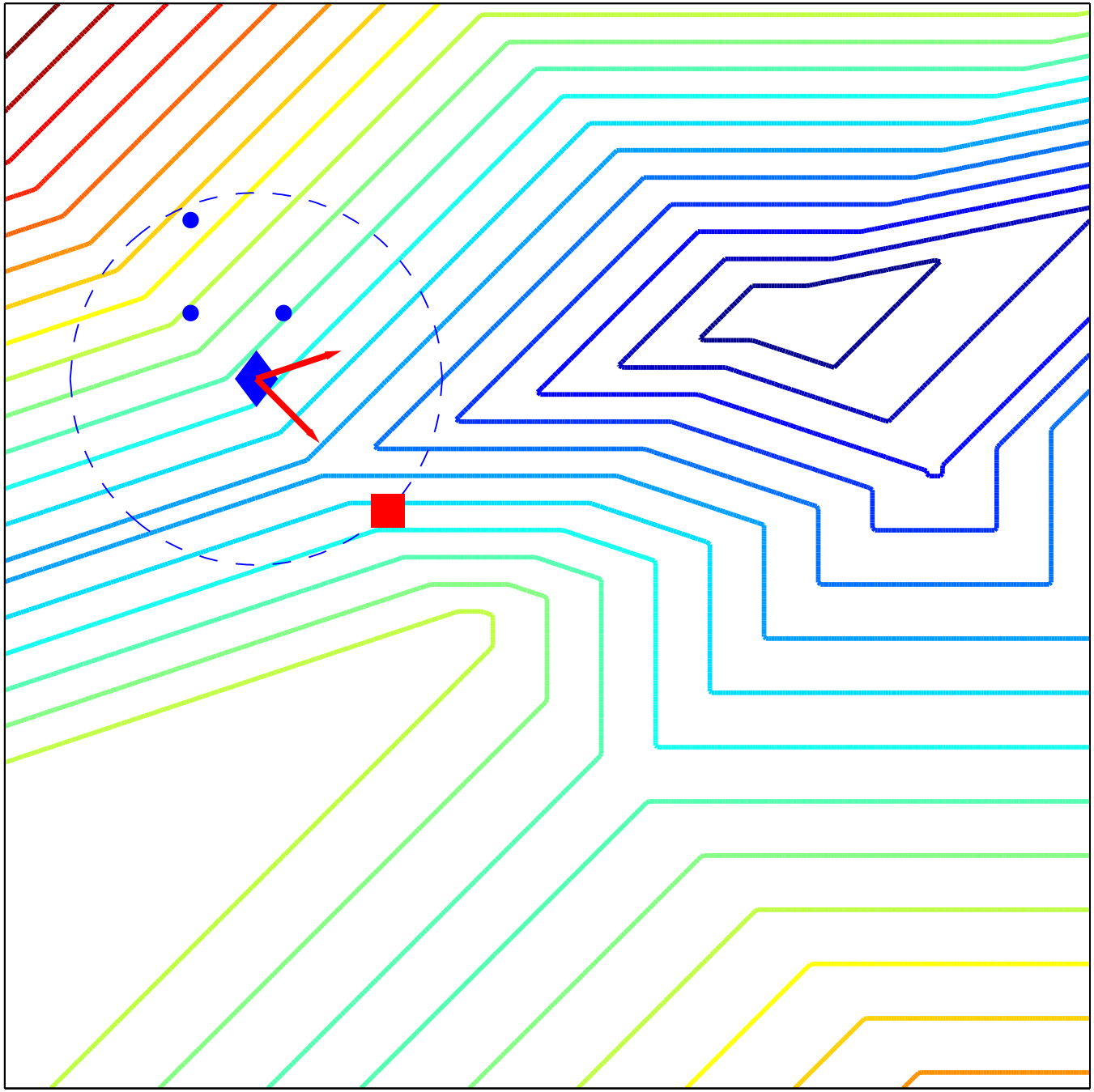


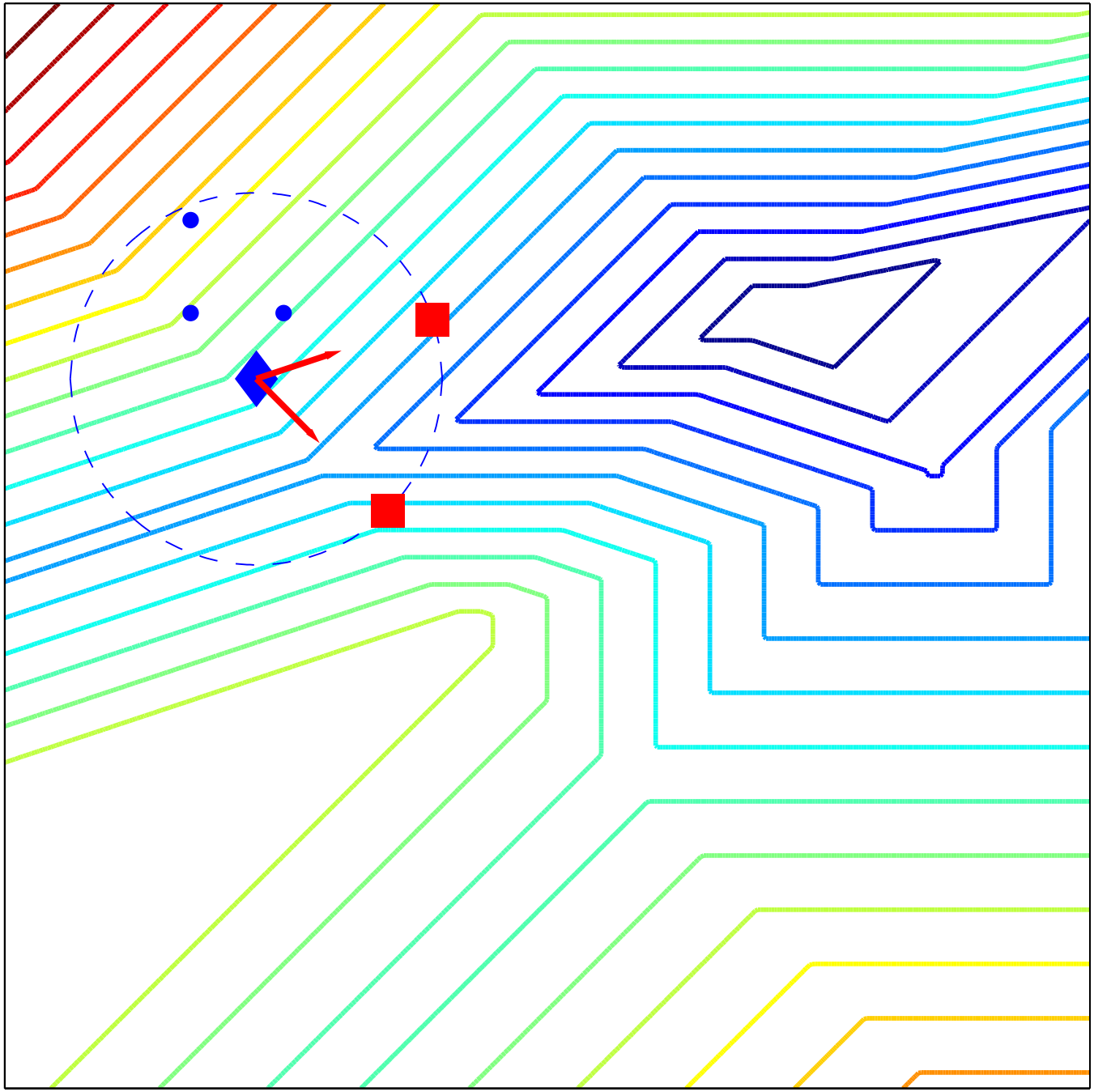


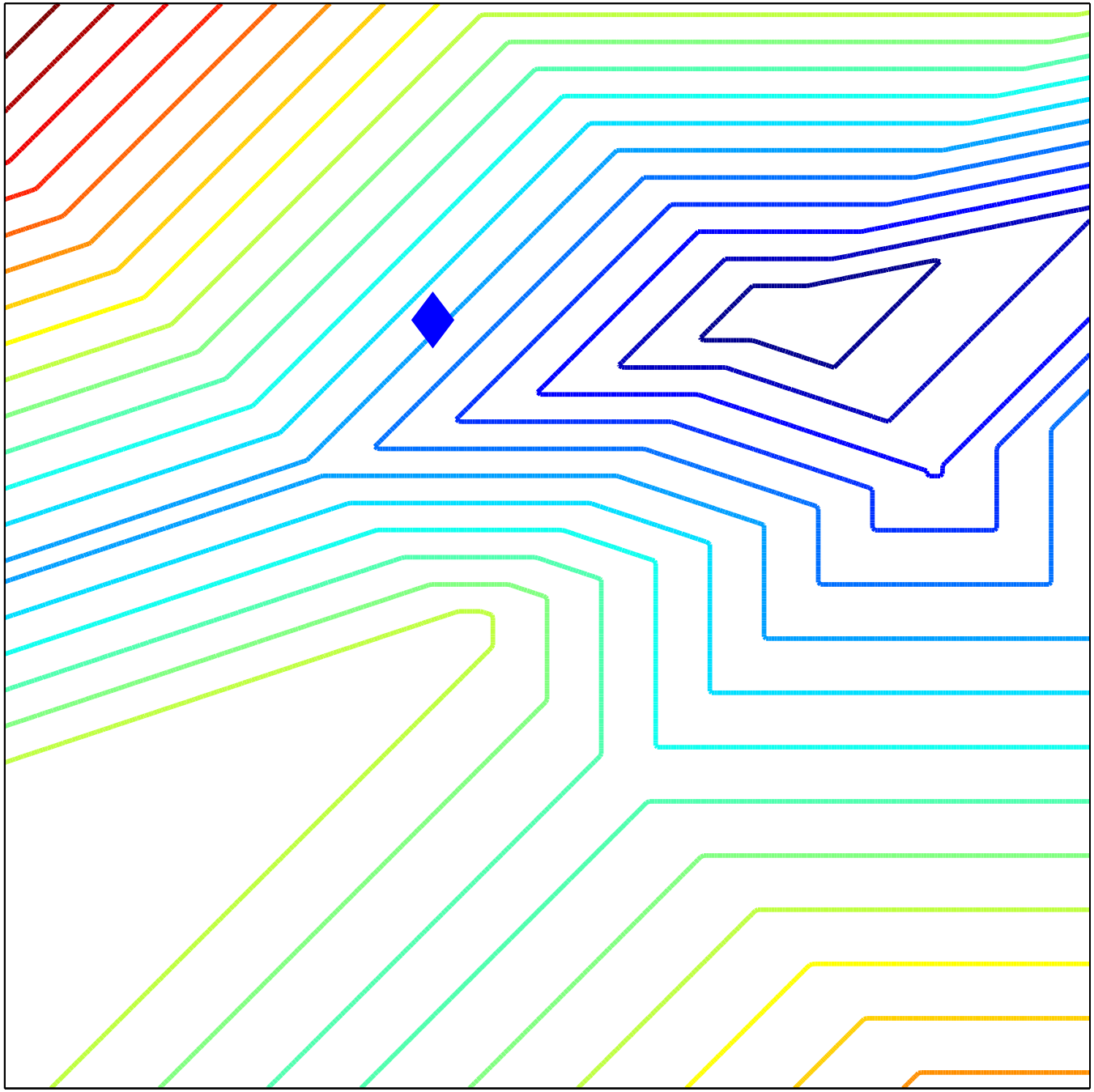


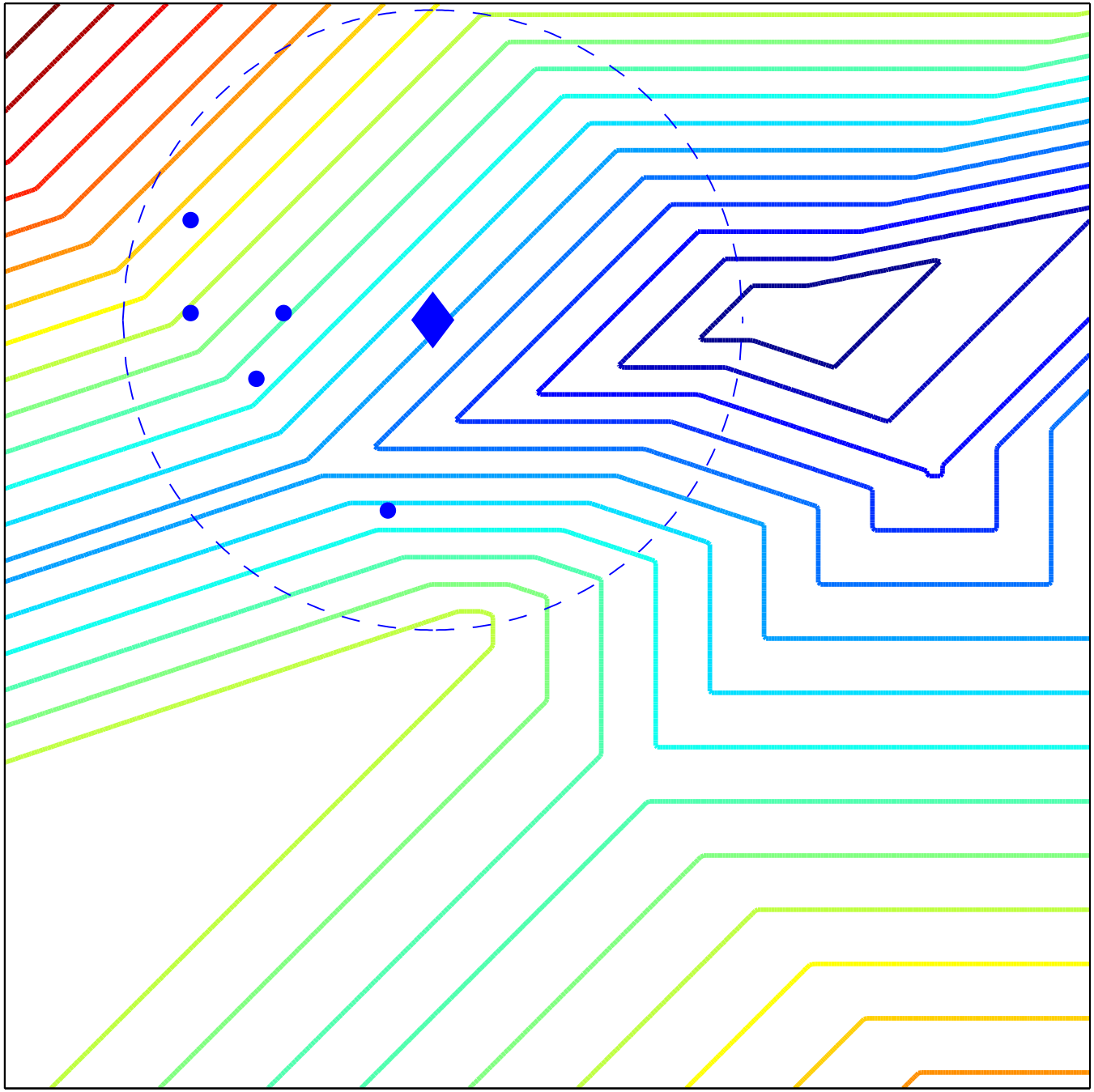


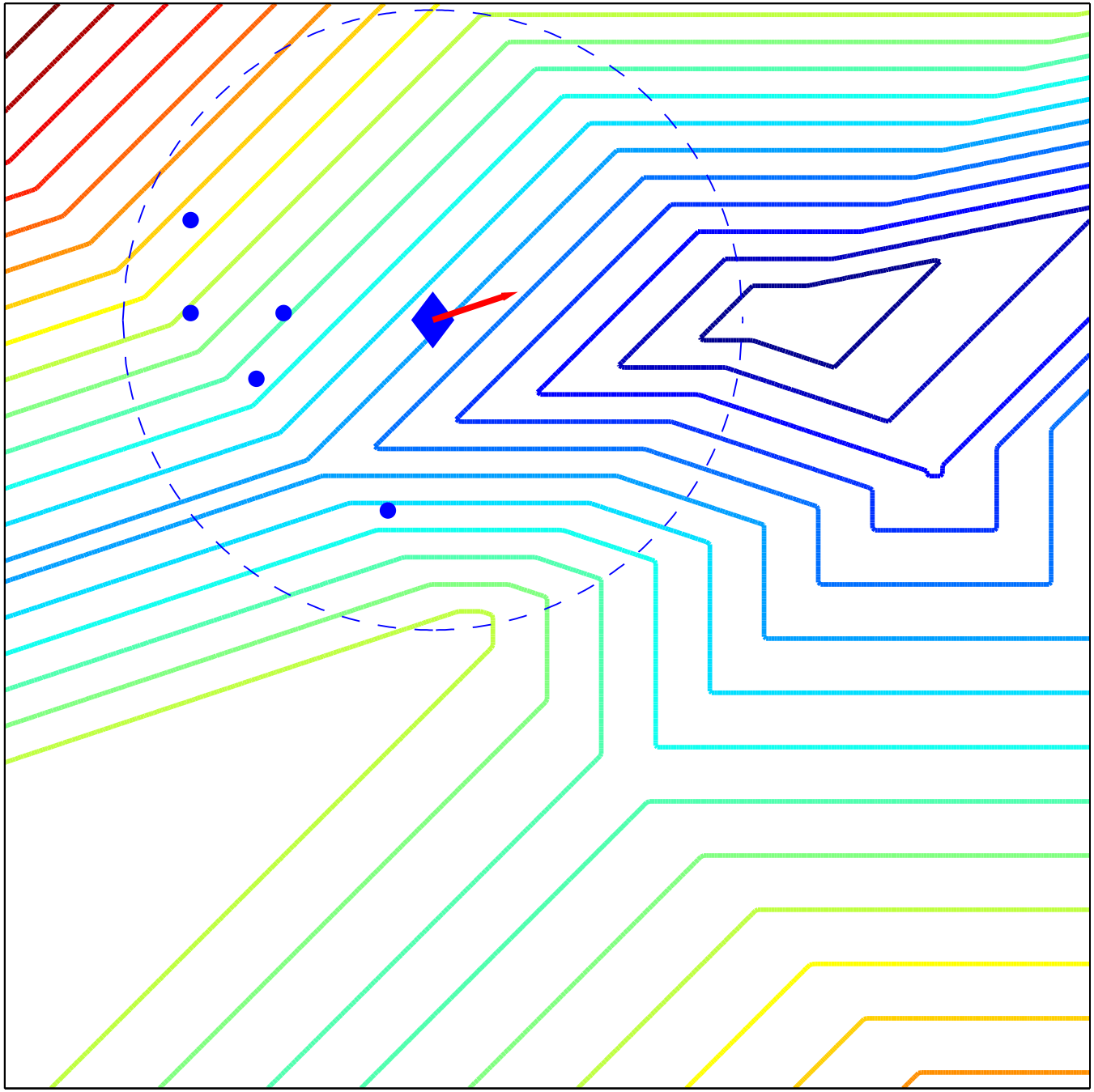


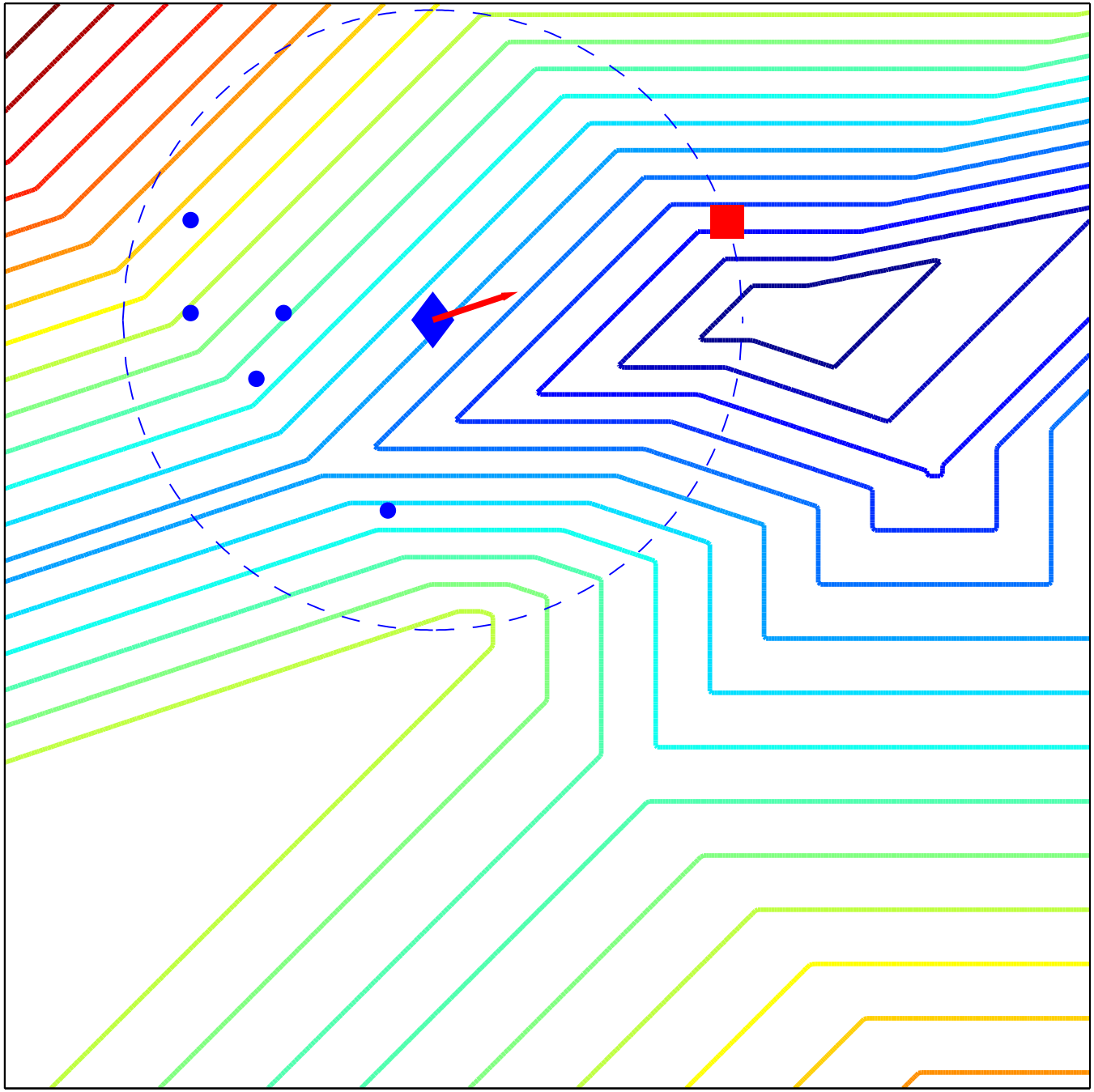


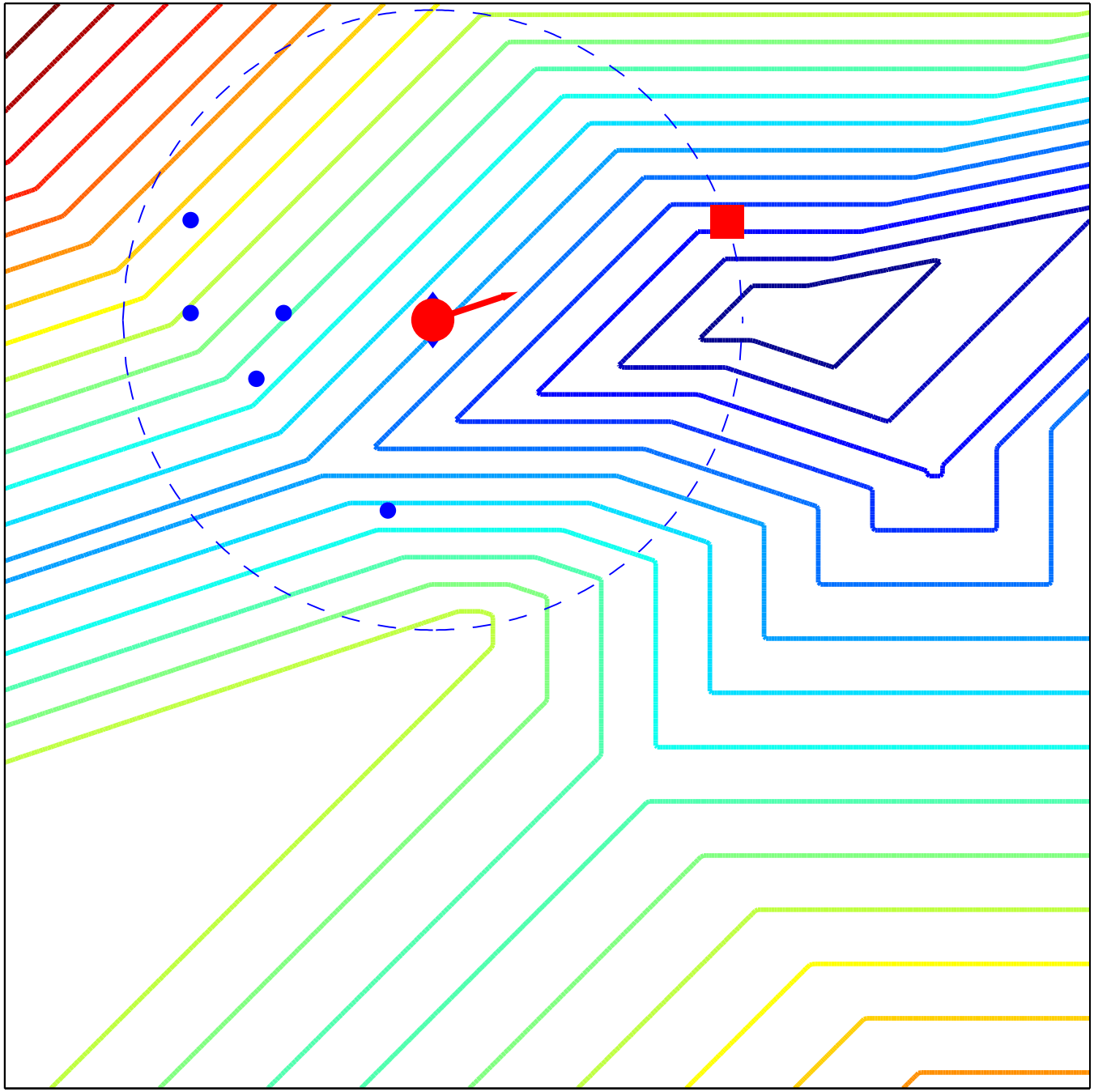


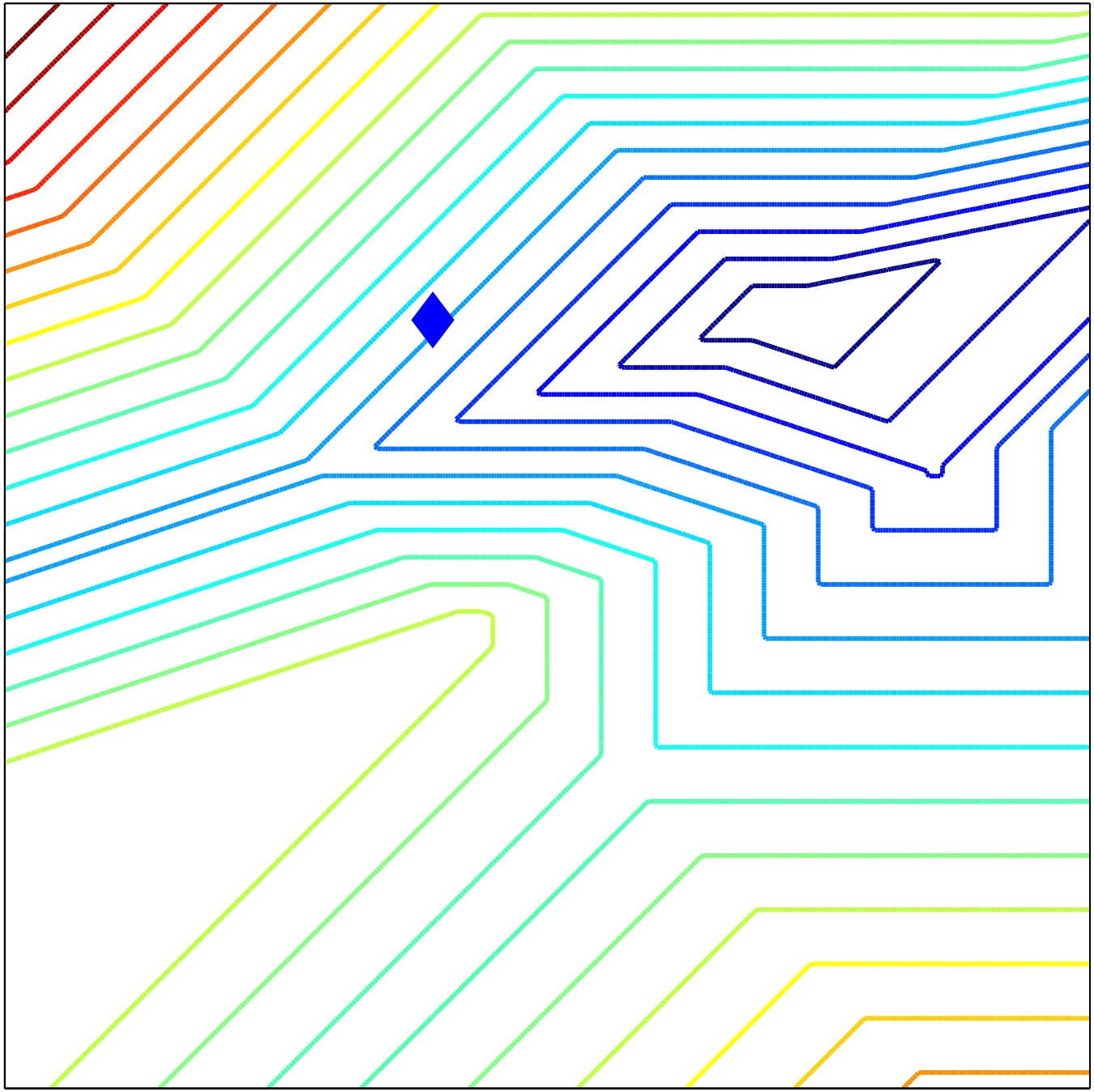


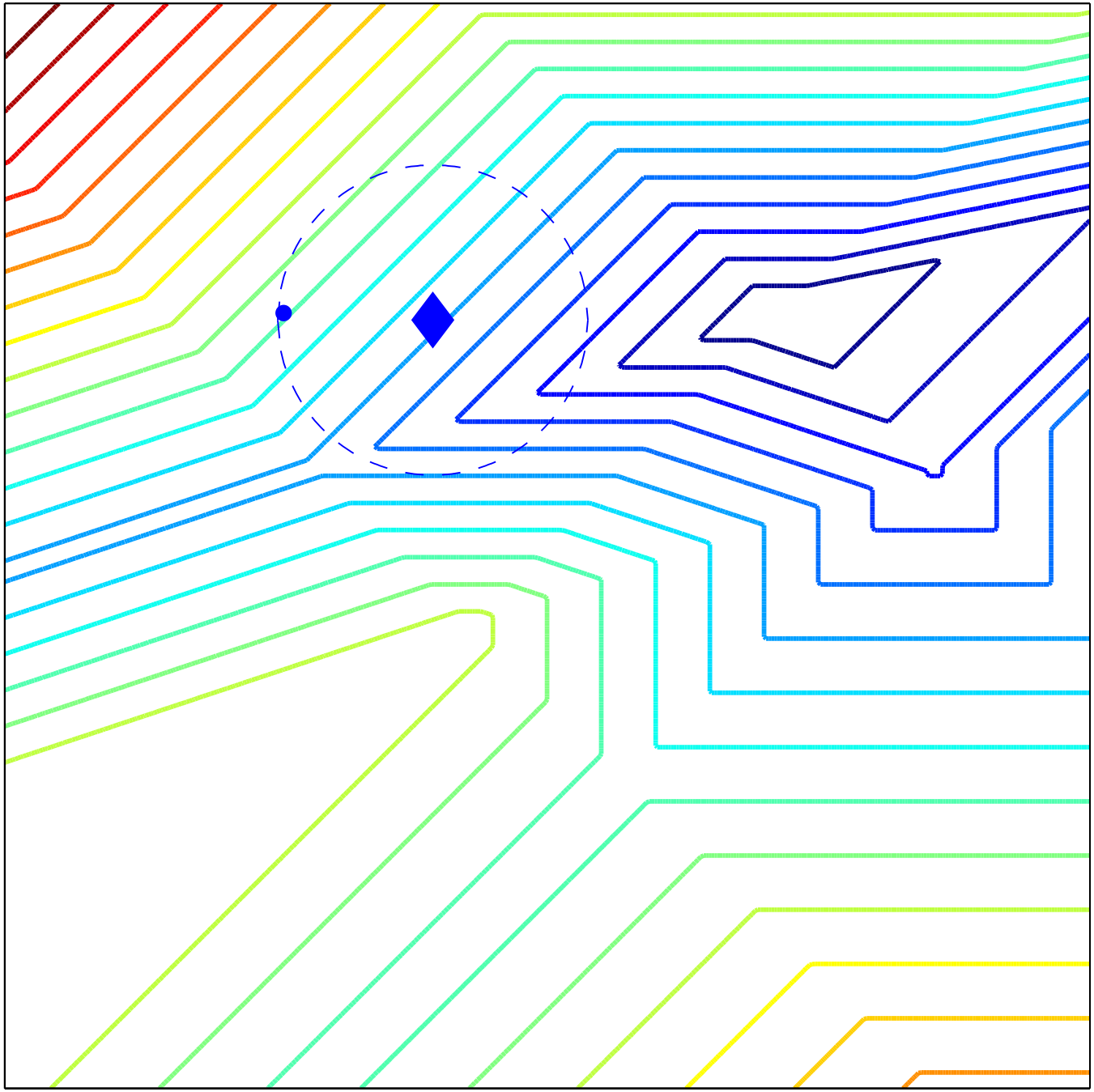


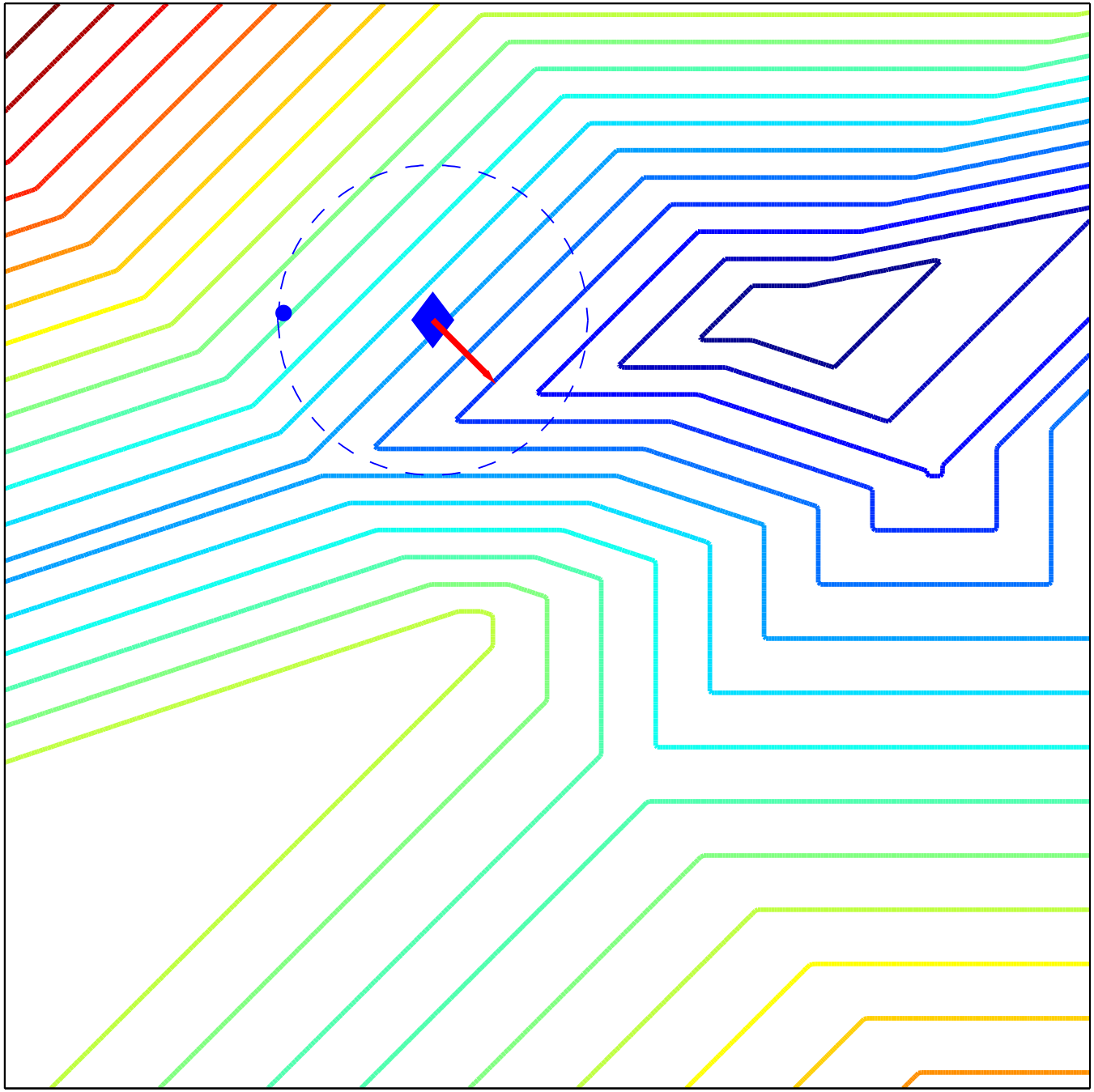


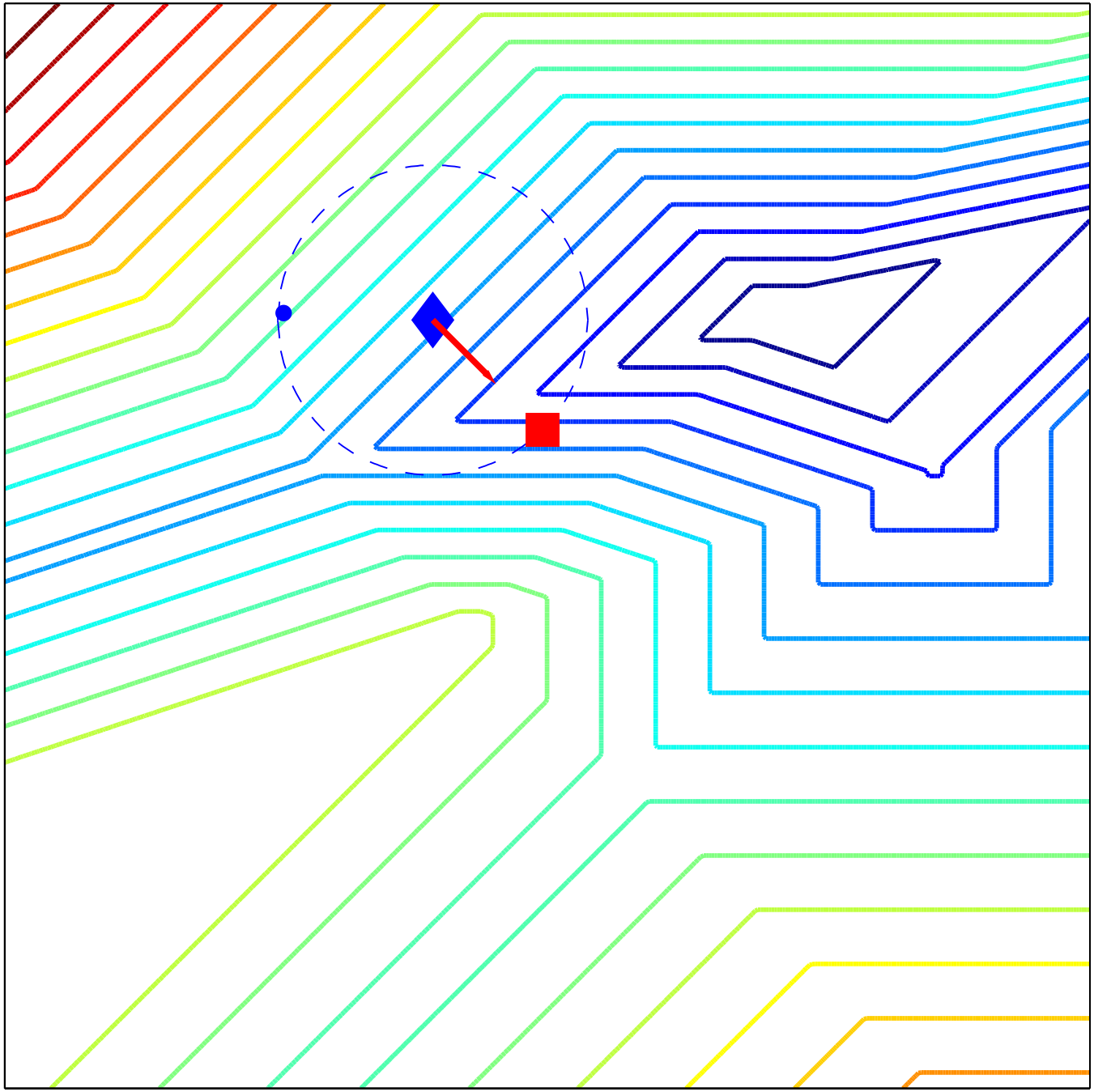


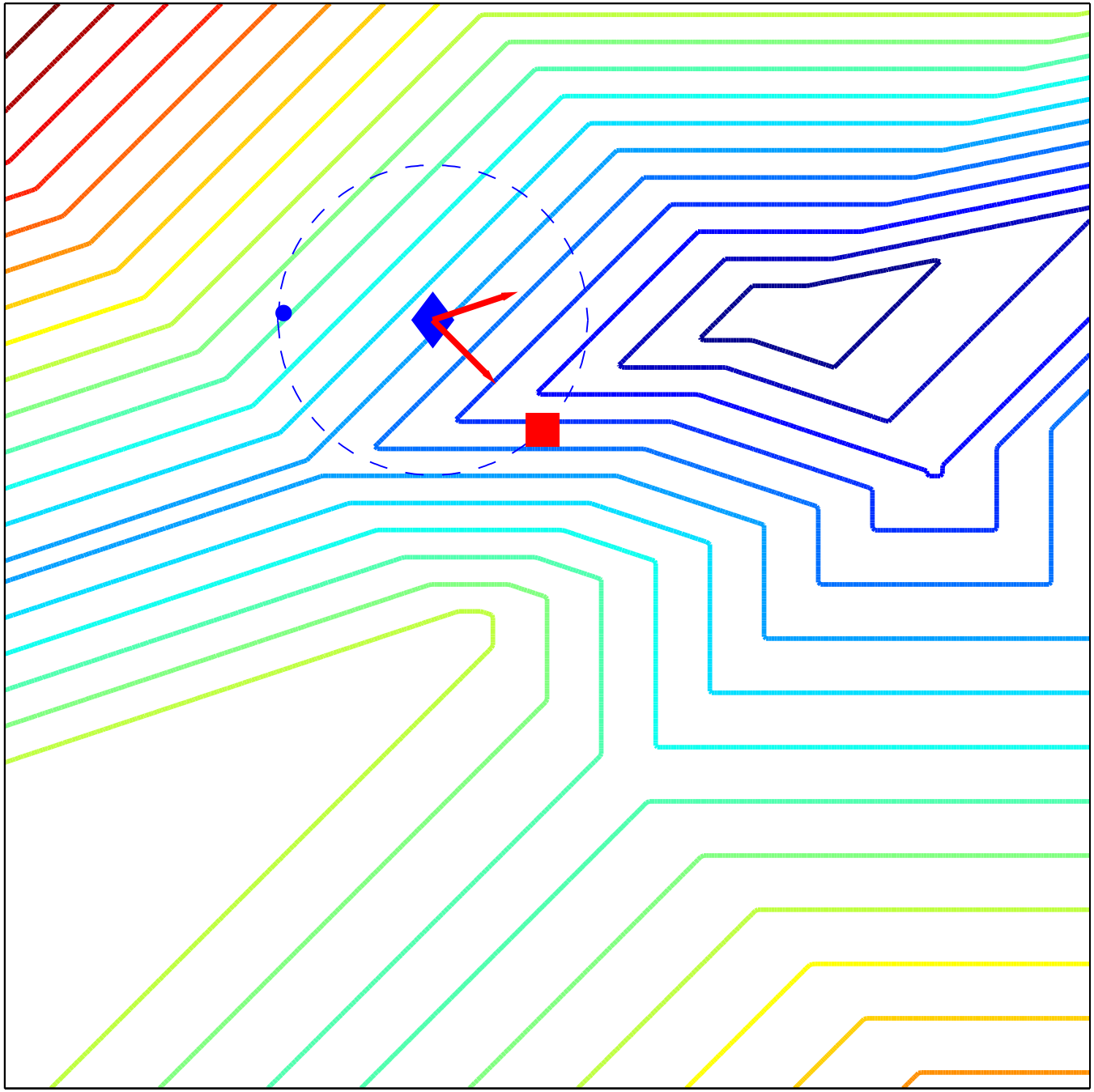


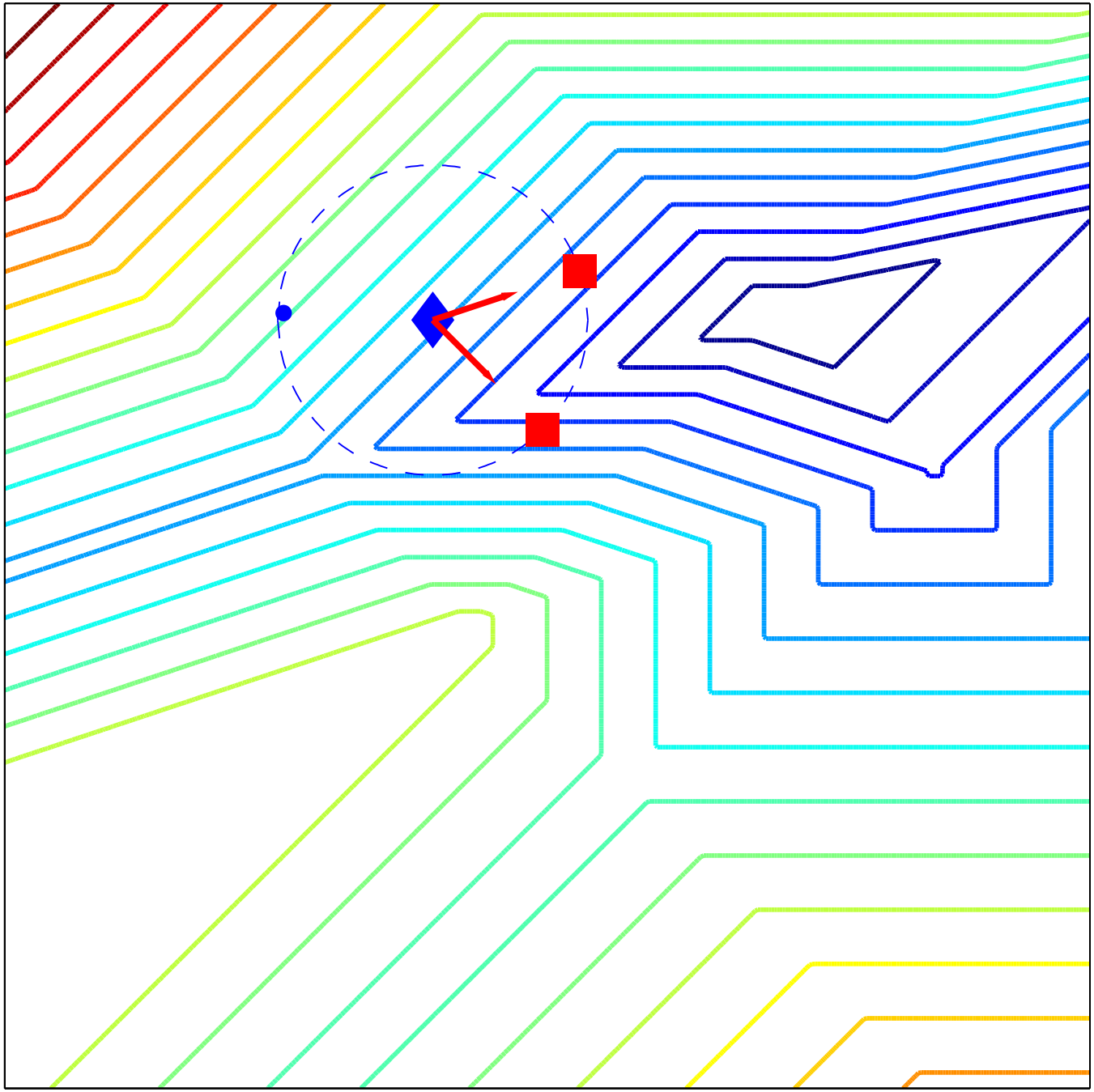


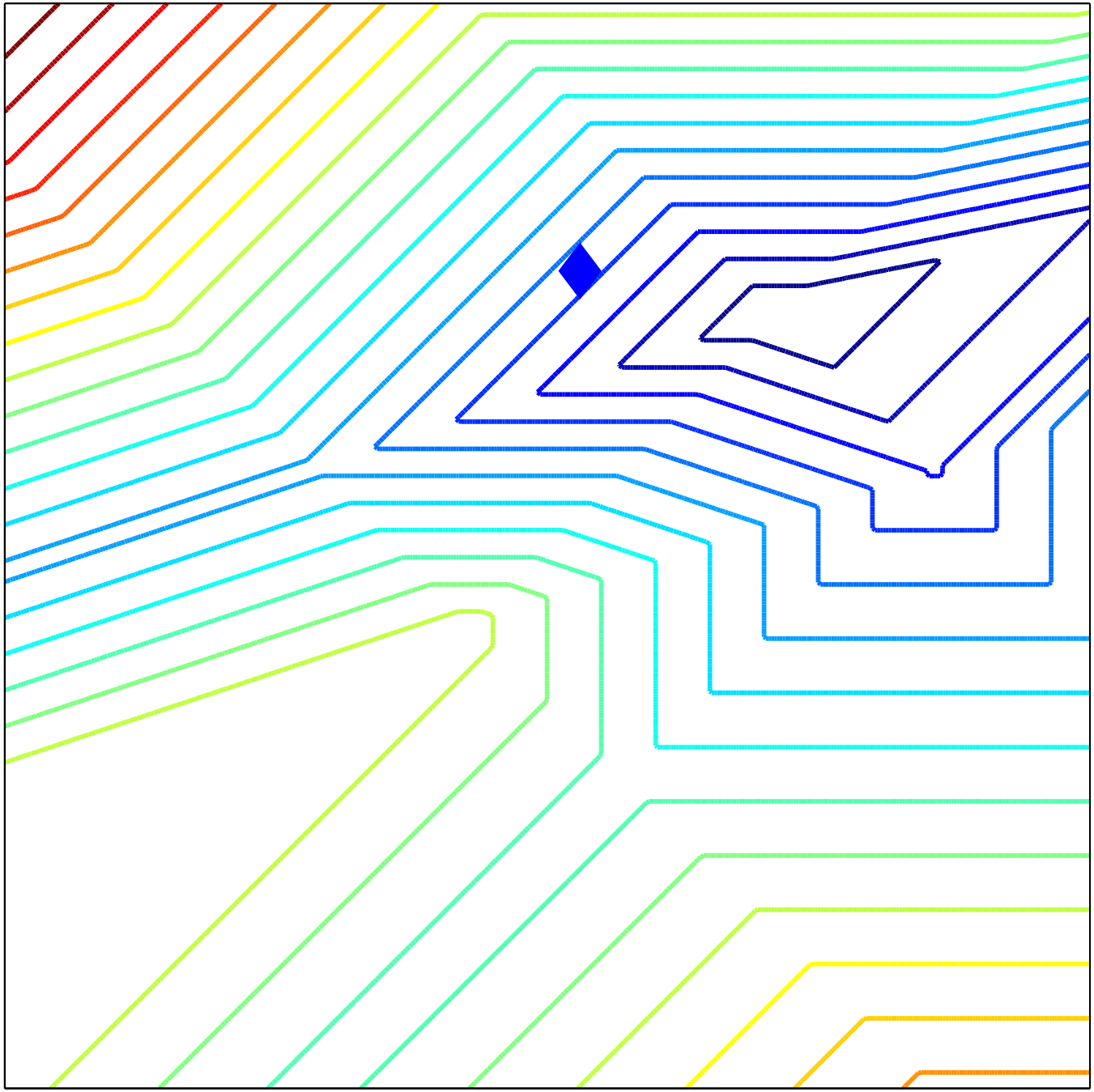


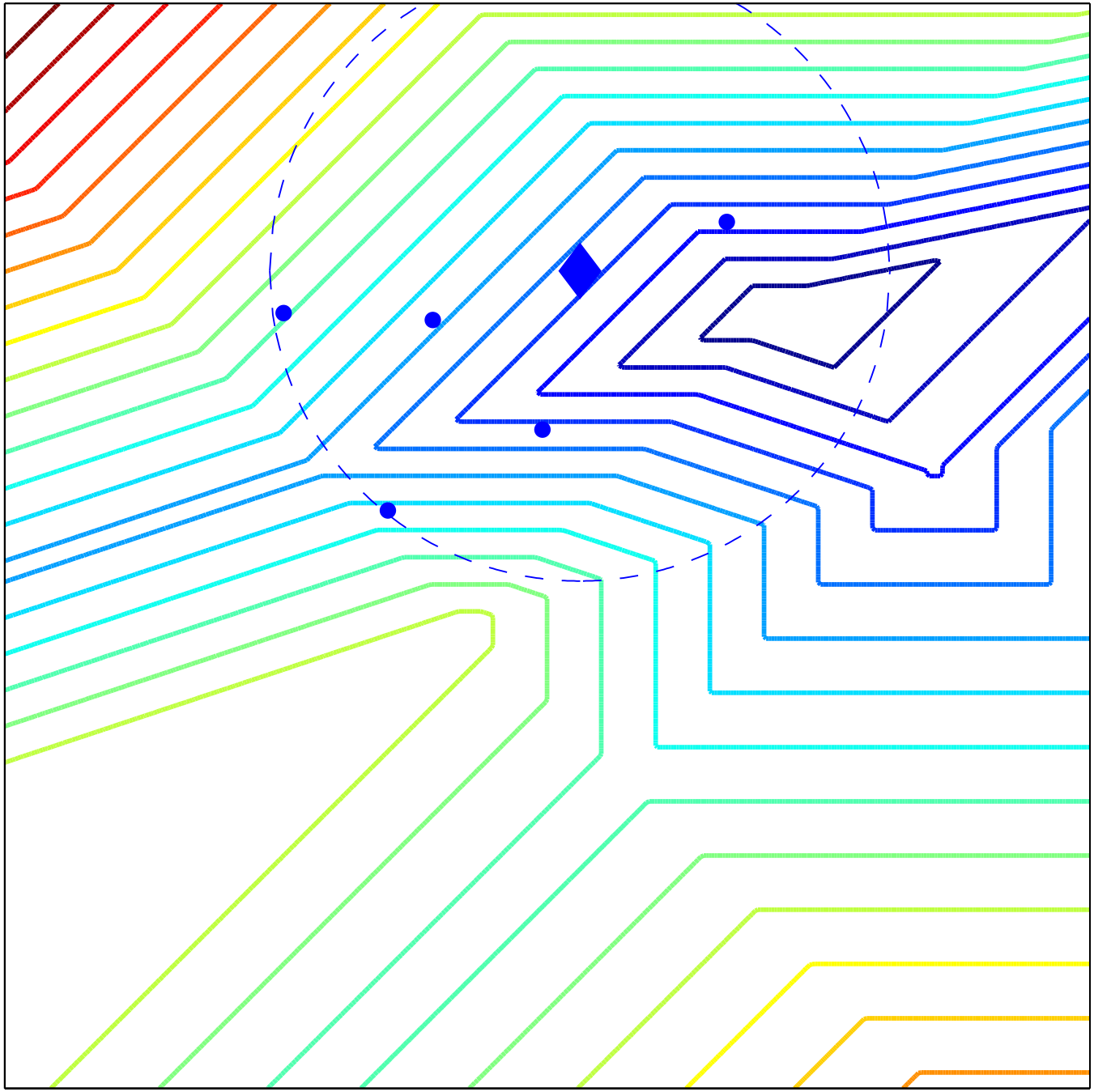


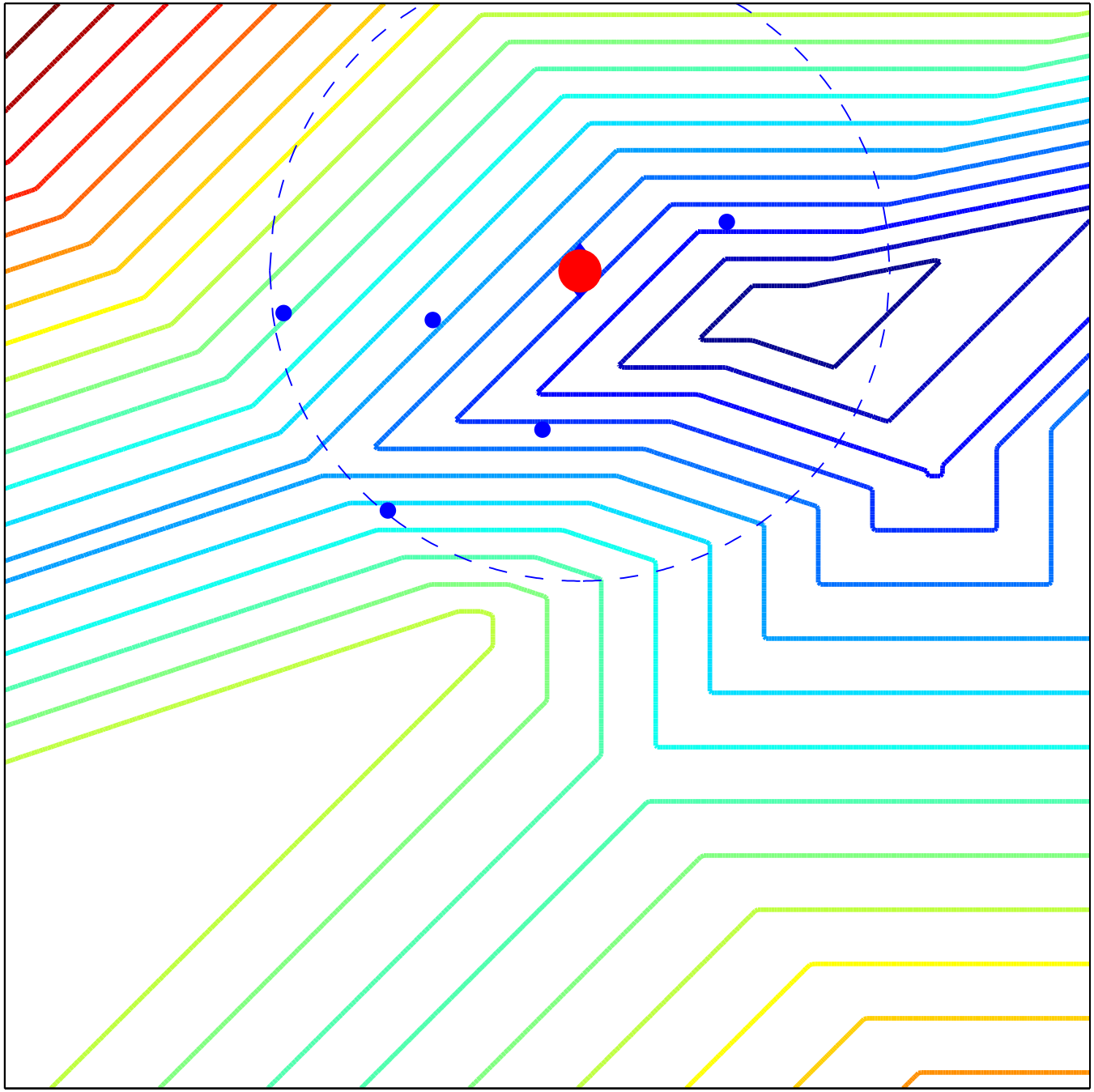


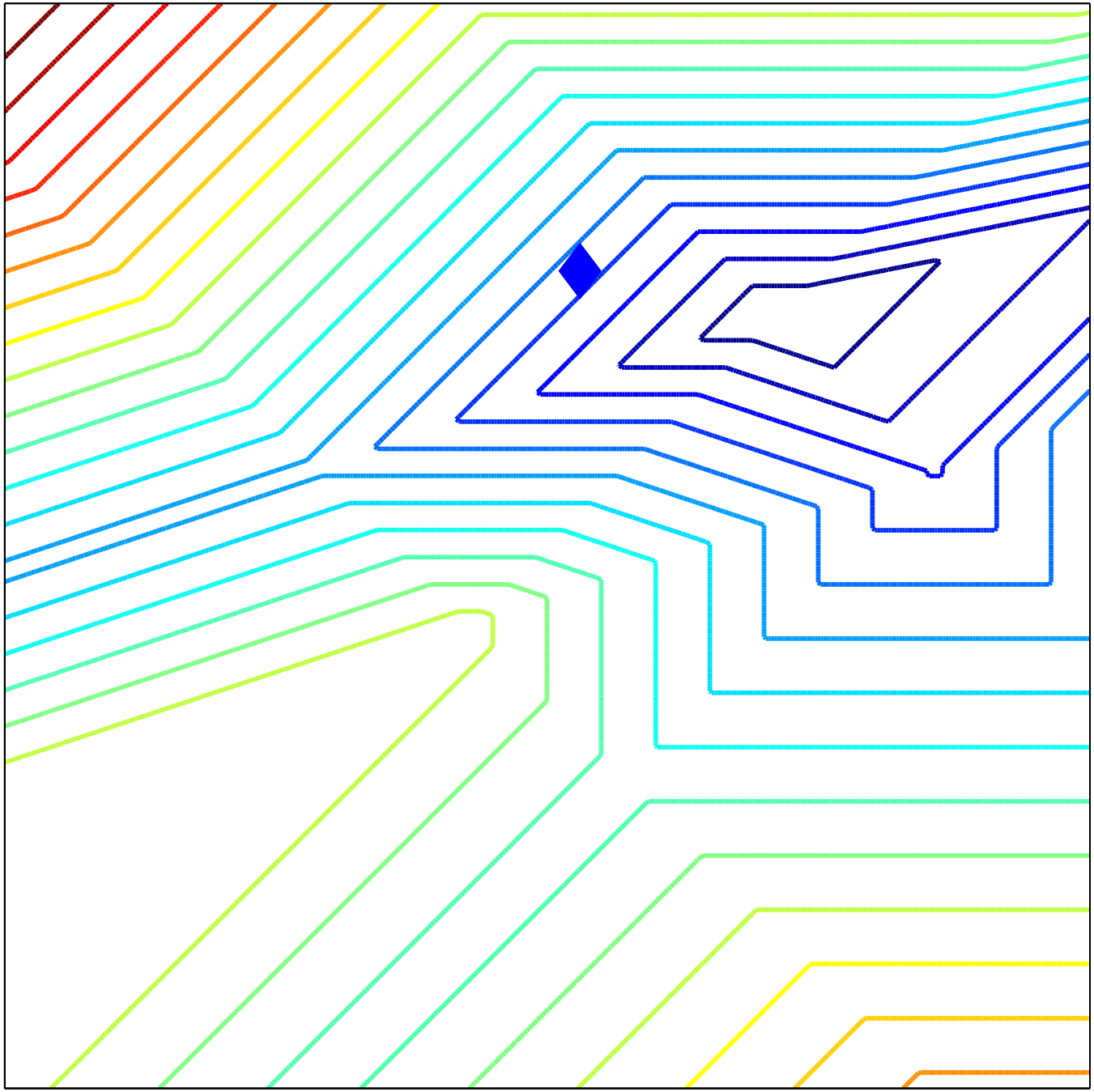


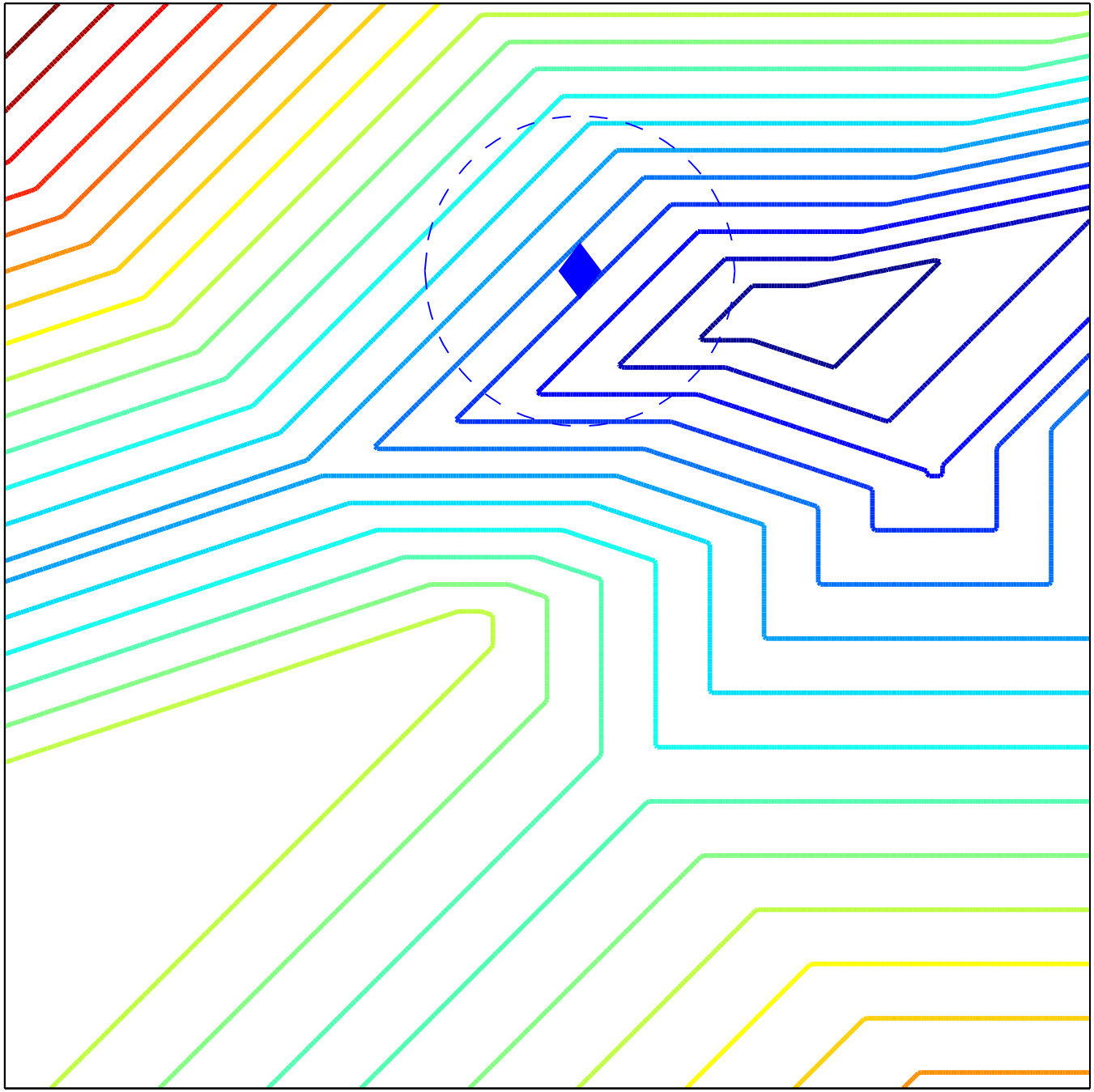


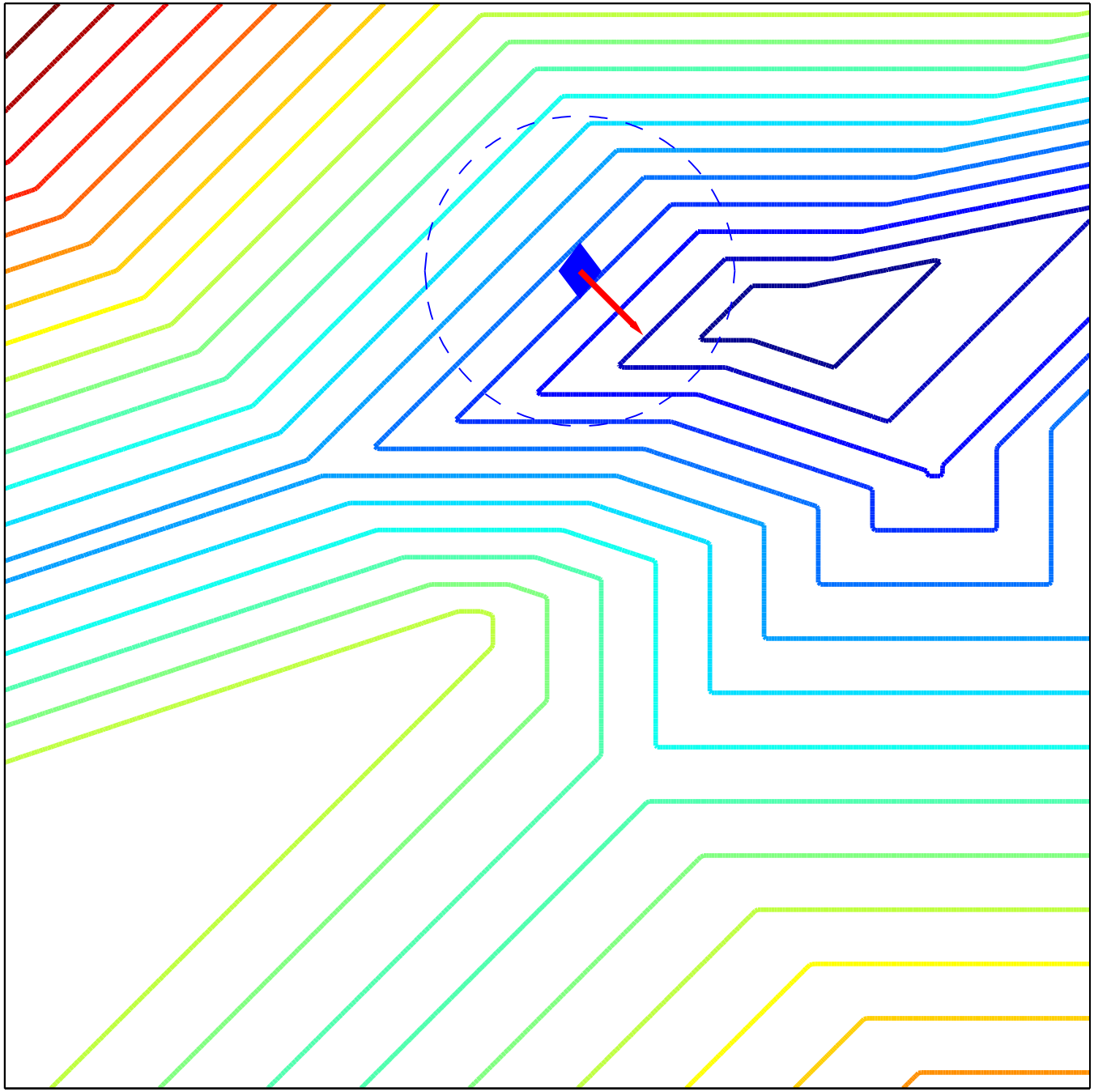


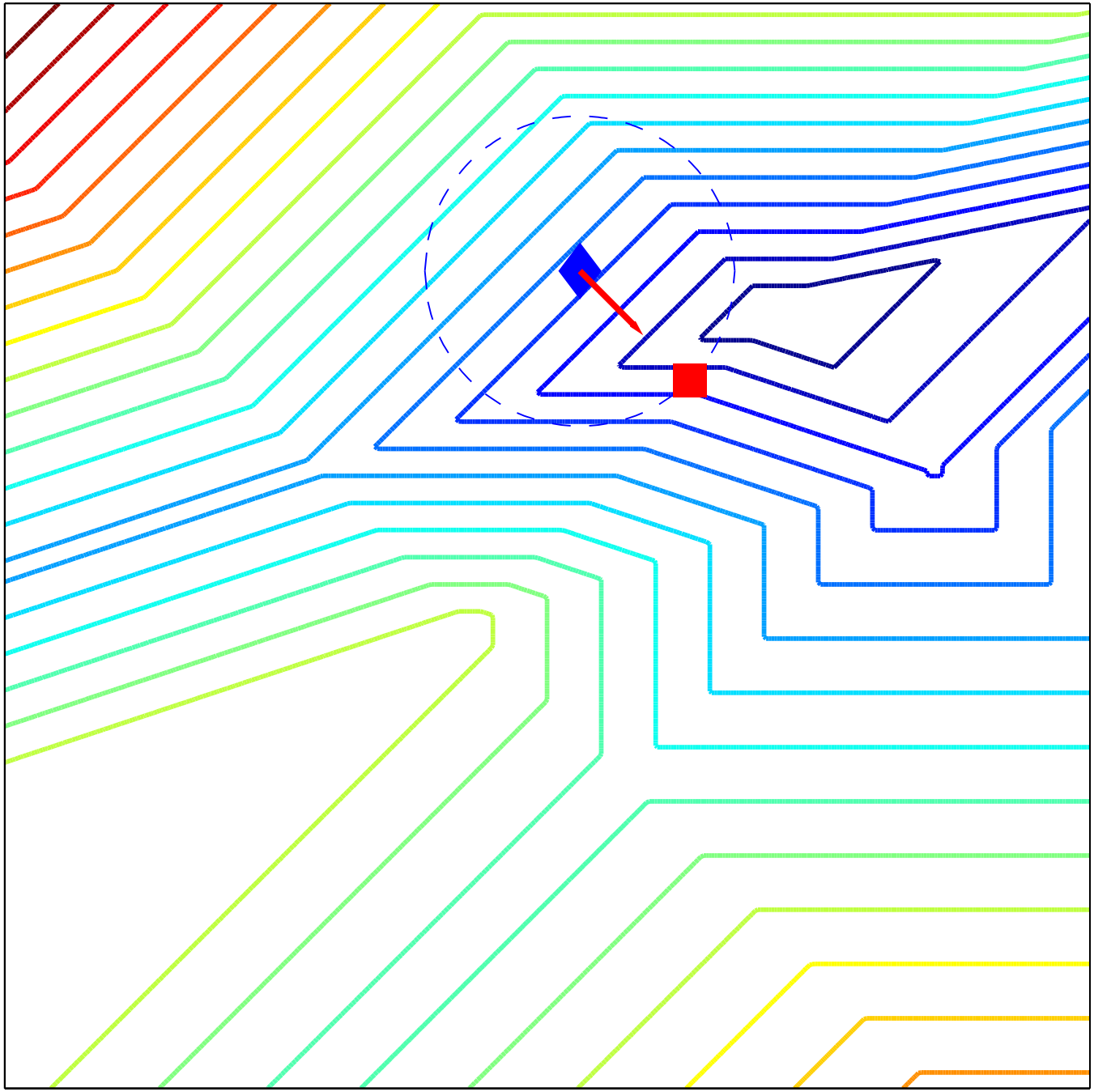


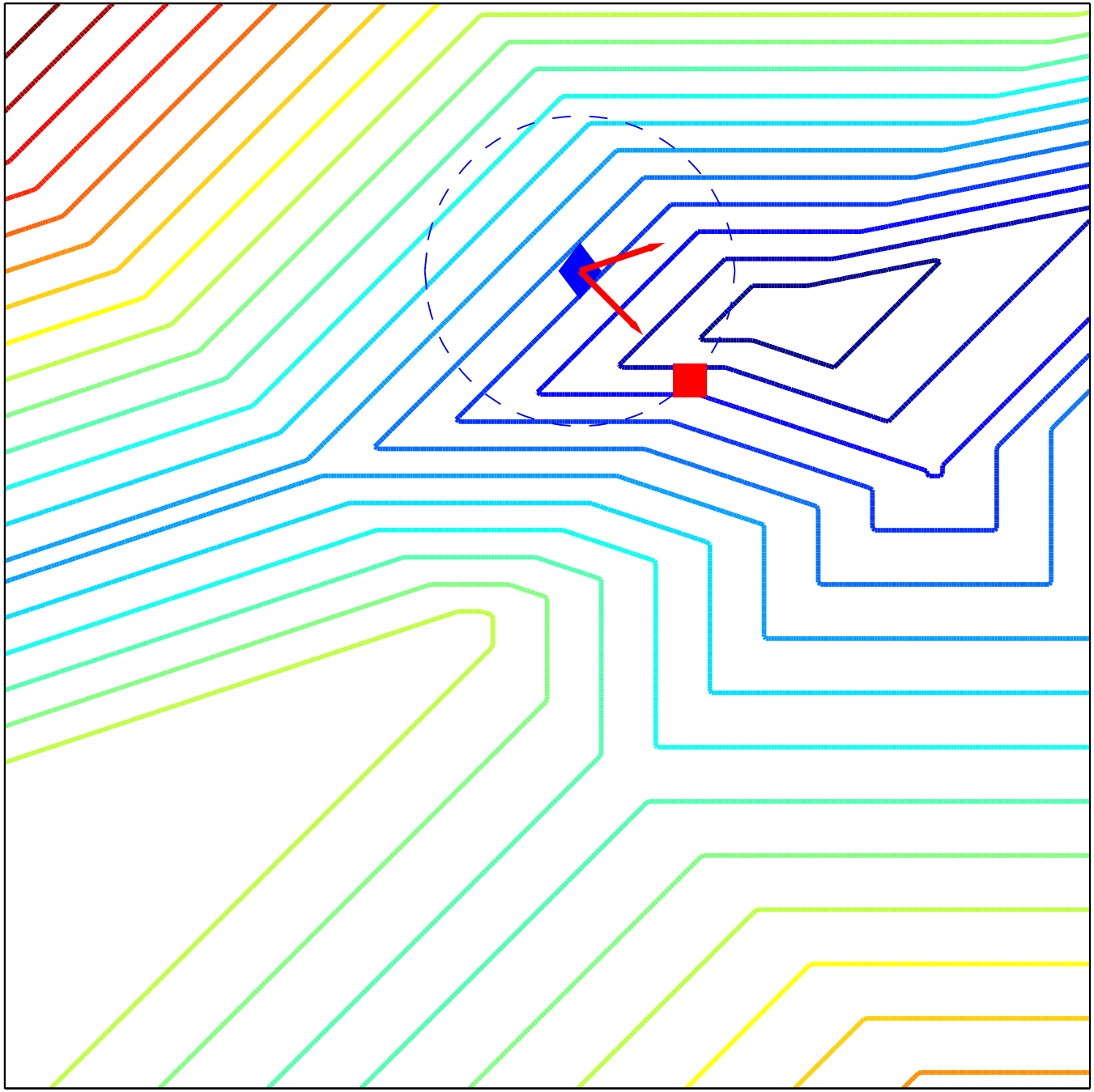


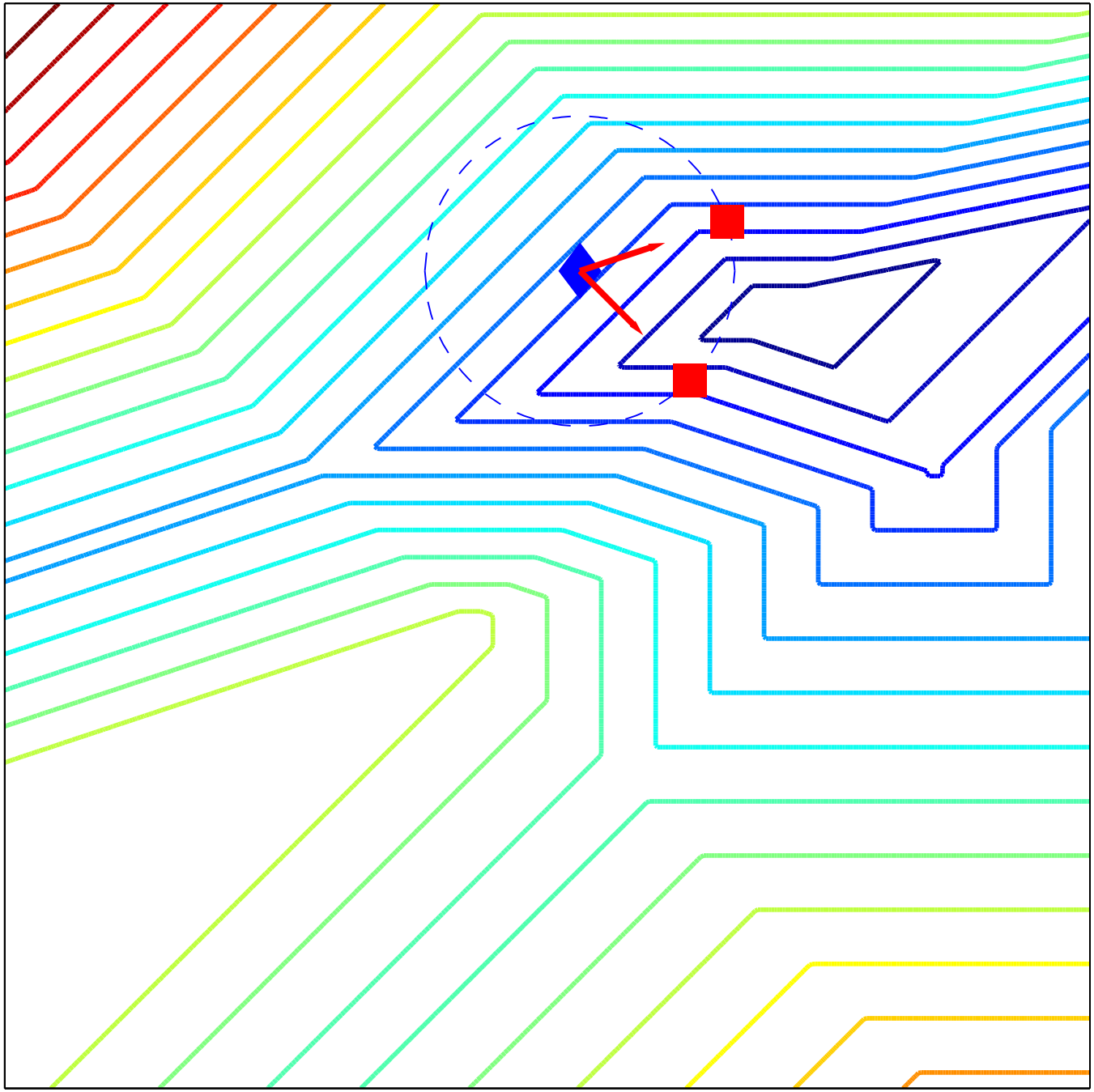


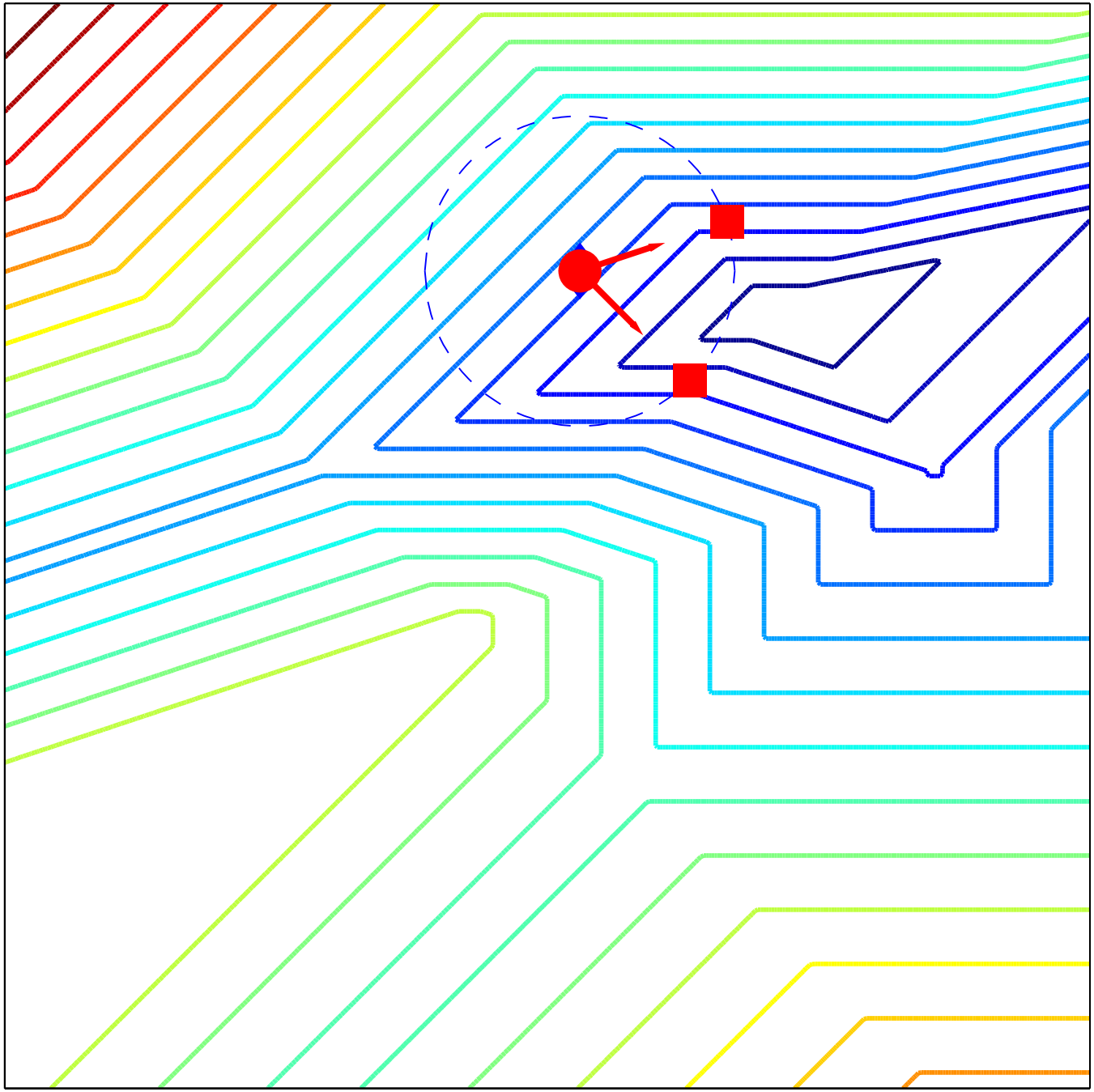












Convergence

- If the trust region radius Δ_k is a sufficiently small multiple of the master model gradient $\left\|g^k\right\|$, the iteration is guaranteed to be successful.
- $\lim_{k\to\infty} \Delta_k = 0$.
- Some subsequence of master model gradients g^k goes zero.
- Zero is in the generalized Clarke subdifferential of cluster points of any subsequence of iterates with master model gradients converging to zero.
- The same holds for cluster points of the sequence of MS4PL iterates.

Test problems

Let h be a censored ℓ_1-loss function. Given data $d \in \mathbb{R}^p$, censors $c \in \mathbb{R}^p$, and the mapping $F : \mathbb{R}^n \to \mathbb{R}^p$, we define

$$f(x) = \sum_{i=1}^{p} |d_i - \max\{F_i(x), c_i\}| .$$

That is, $\psi = 0$, and

$$h(y) = \sum_{i=1}^{p} |d_i - \max\{y_i, c_i\}| .$$

Test problems

Let h be a censored ℓ_1-loss function. Given data $d \in \mathbb{R}^p$, censors $c \in \mathbb{R}^p$, and the mapping $F : \mathbb{R}^n \to \mathbb{R}^p$, we define

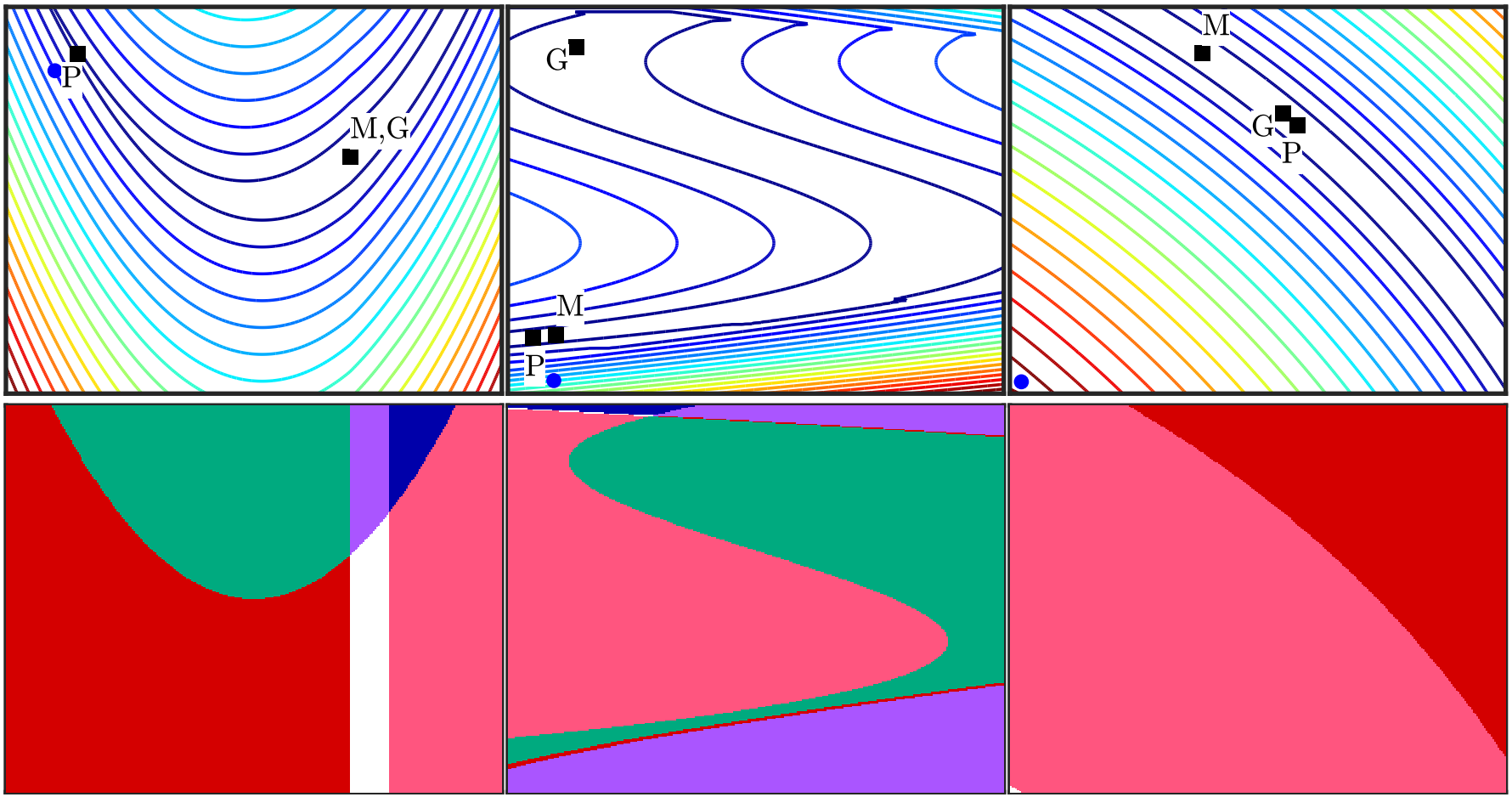
$$f(x) = \sum_{i=1}^{p} |d_i - \max\{F_i(x), c_i\}| .$$

That is, $\psi = 0$, and

$$h(y) = \sum_{i=1}^{p} |d_i - \max\{y_i, c_i\}| .$$

Define F to be the 53 vector mapping in the Móre and Wild benchmarking set. $2 \leq n \leq 12$, $2 \leq p \leq 45$.

Examples



Algorithms to compare

MS4PL-1 Using manifolds at x^k [F, knowledge of h]

MS4PL-2 Using manifolds in $\mathcal{B}(x^k, \Delta_k)$ [F, knowledge of h]

Algorithms to compare

MS4PL-1 Using manifolds at x^k [F, knowledge of h]

MS4PL-2 Using manifolds in $\mathcal{B}(x^k, \Delta_k)$ [F, knowledge of h]

PLC POUNDERs using a single manifold active at x^k to form a master model [F, single element of $\partial_B h$]

Algorithms to compare

MS4PL-1 Using manifolds at x^k [F, knowledge of h]

MS4PL-2 Using manifolds in $\mathcal{B}(x^k, \Delta_k)$ [F, knowledge of h]

PLC POUNDERs using a single manifold active at x^k to form a master model [F, single element of $\partial_B h$]

SLQP-GS Gradient sampling algorithm [Curtis] [f, $\partial_B f$ (via ∇F)]

GRANSO BFGS-SQP [Mitchell, Curtis, Overton.] [f, $\partial_B f$ (via ∇F)]

Algorithms to compare

MS4PL-1 Using manifolds at x^k [F, knowledge of h]

MS4PL-2 Using manifolds in $\mathcal{B}(x^k, \Delta_k)$ [F, knowledge of h]

PLC POUNDERs using a single manifold active at x^k to form a master model [F, single element of $\partial_B h$]

SLQP-GS Gradient sampling algorithm [Curtis] [f, $\partial_B f$ (via ∇F)]

GRANSO BFGS-SQP [Mitchell, Curtis, Overton.] [f, $\partial_B f$ (via ∇F)]

MS4PL-1-grad Using manifolds at x^k [F, knowledge of h, ∇F for models]

Tests

f test A method s solves a problem p to a level τ after j function evaluations if

$$f(x^0) - f(x^j) \geq (1 - \tau)(f(x^0) - \tilde{f}_p)$$

x^0 is the problem's starting point, and $\tilde{f}_p$ is the best-found function value.

Tests

f test A method s solves a problem p to a level τ after j function evaluations if

$$f(x^0) - f(x^j) \geq (1 - \tau)(f(x^0) - \tilde{f}_p)$$

x^0 is the problem's starting point, and $\tilde{f}_p$ is the best-found function value.

$\partial_C f$ test Sample gradients.

Tests

f test A method s solves a problem p to a level τ after j function evaluations if

$$f(x^0) - f(x^j) \geq (1 - \tau)(f(x^0) - \tilde{f}_p)$$

x^0 is the problem's starting point, and $\tilde{f}_p$ is the best-found function value.

$\partial_C f$ test Sample gradients.

Draw 30 points uniformly from $B(x^j, 10^{-8})$ for each point x^j evaluated by each method.

Tests

f test A method s solves a problem p to a level τ after j function evaluations if

$$f(x^0) - f(x^j) \geq (1 - \tau)(f(x^0) - \tilde{f}_p)$$

x^0 is the problem's starting point, and $\tilde{f}_p$ is the best-found function value.

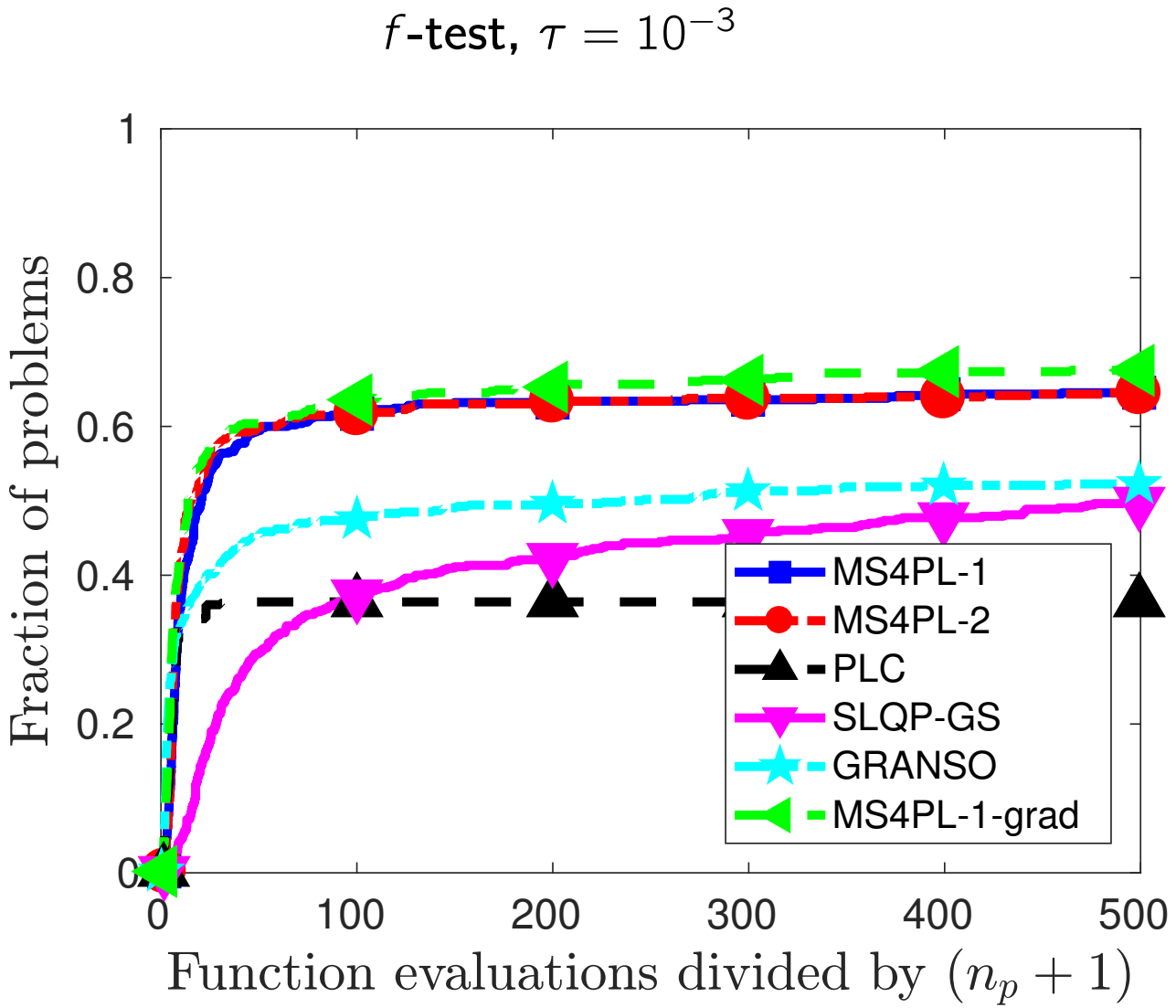
$\partial_C f$ test Sample gradients.

Draw 30 points uniformly from $B(x^j, 10^{-8})$ for each point x^j evaluated by each method.

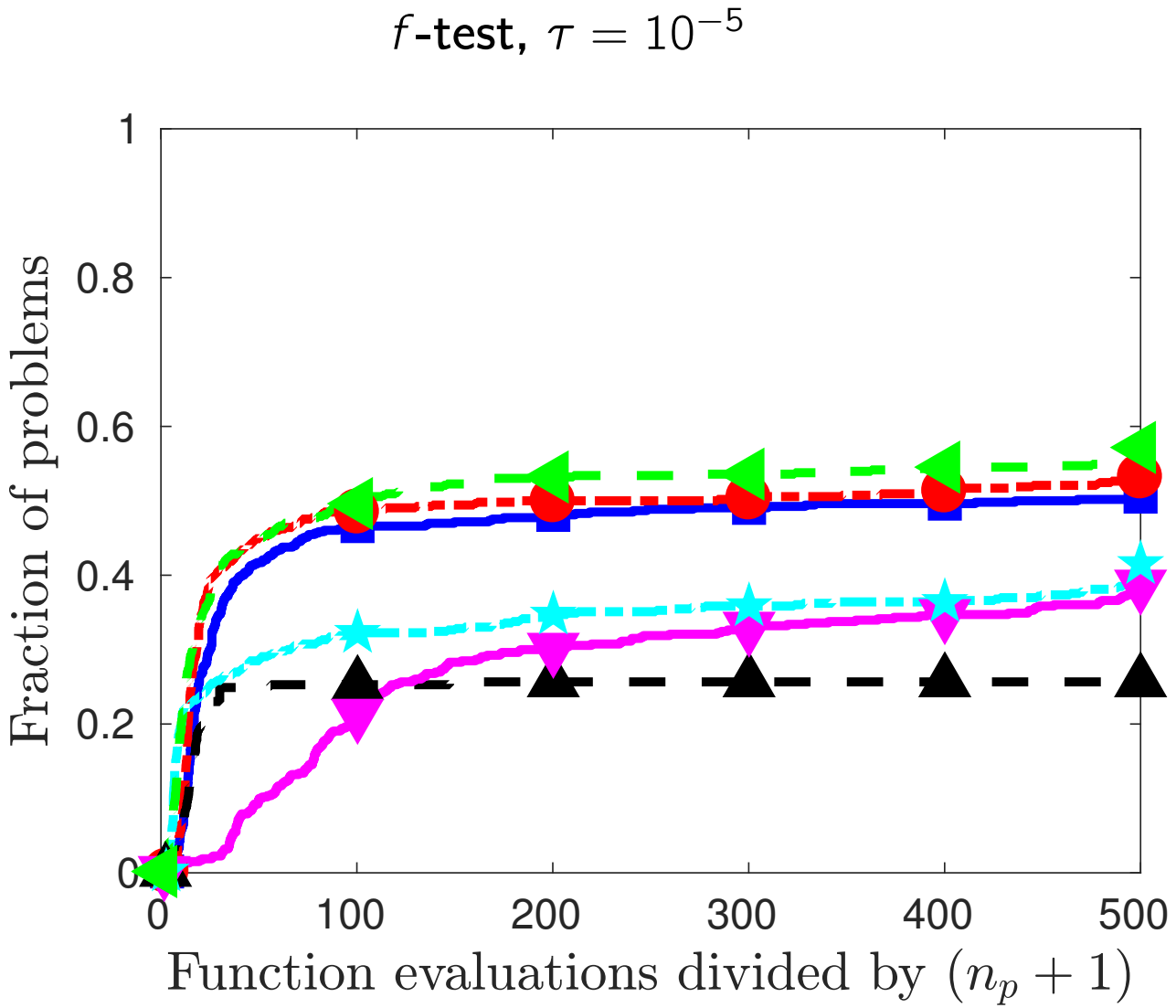
s solves p to a level τ after j function evaluations if

$$\left\|\tilde{g}^j\right\| \leq \tau \left\|\tilde{g}^0\right\|$$

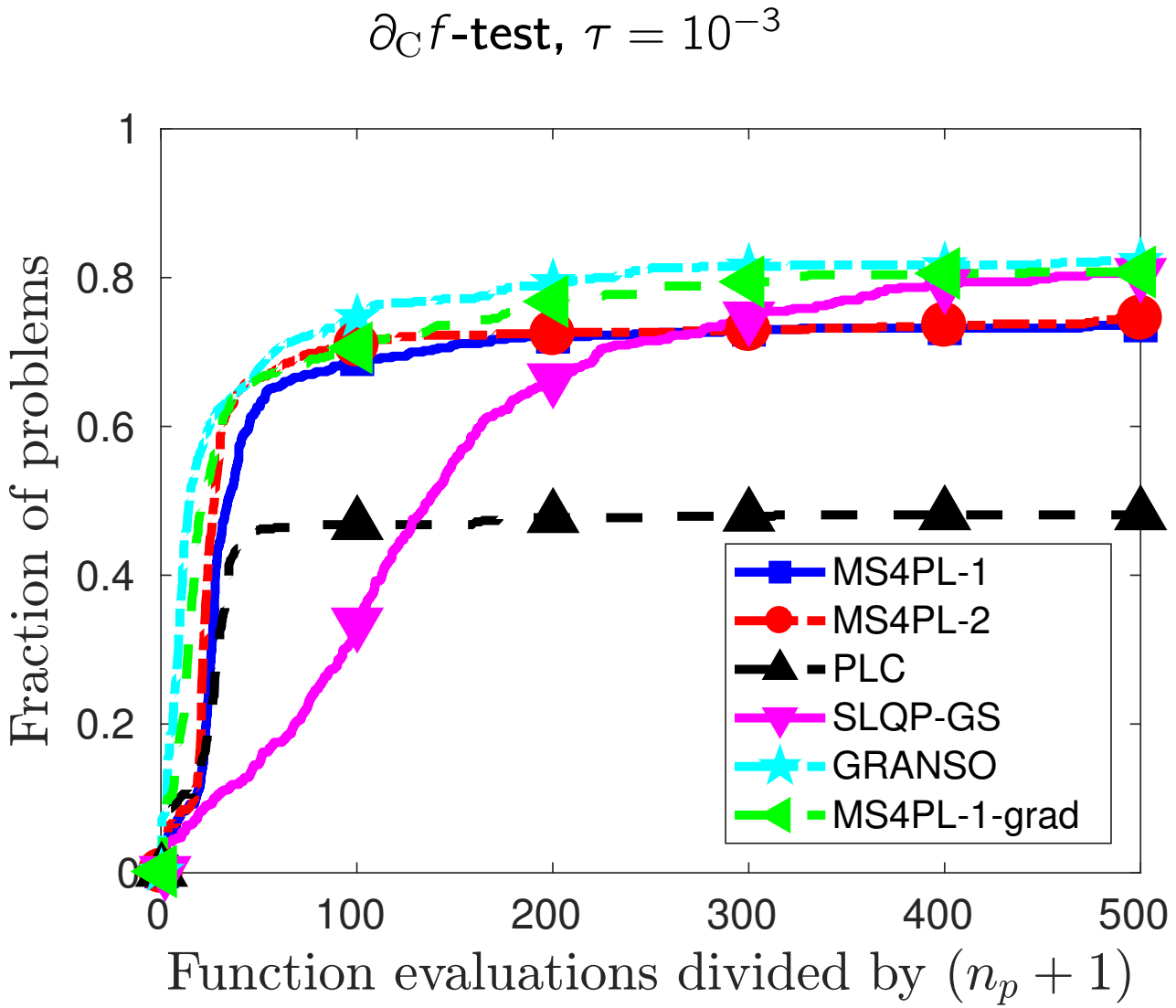
Data profiles



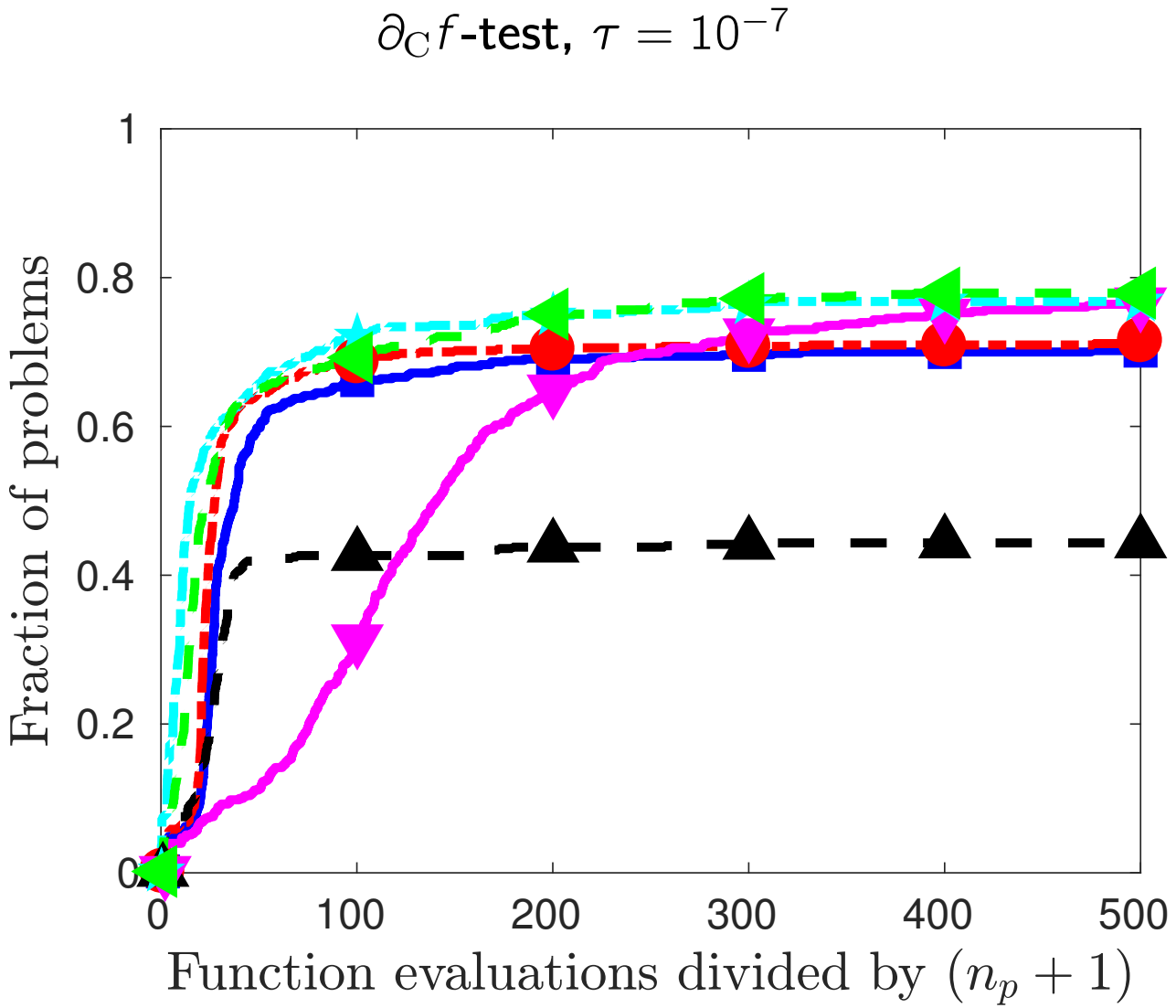
Data profiles



Data profiles



Data profiles



Conclusions

When optimizing functions of the form $h(F(x))$ when

- h is "easy"
- F is "hard"

it can be advantageous to model F_i and then combine those models via known information about h.

Conclusions

When optimizing functions of the form $h(F(x))$ when

- h is "easy"
- F is "hard"

it can be advantageous to model F_i and then combine those models via known information about h.

Email `jmlarson@anl.gov` for a preprint.

Thank you!